

# Computation- and Annotation- Efficient Deep Learning for Biomedical Image Analysis

Lead Guest Editor: Jialin Peng

Guest Editors: Dong Nie and Peijun Hu





---

# **Computation- and Annotation-Efficient Deep Learning for Biomedical Image Analysis**



Journal of Healthcare Engineering

---

**Computation- and Annotation-Efficient  
Deep Learning for Biomedical Image  
Analysis**

Lead Guest Editor: Jialin Peng

Guest Editors: Dong Nie and Peijun Hu



---


Copyright © 2021 Hindawi Limited. All rights reserved.

This is a special issue published in "Journal of Healthcare Engineering." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Associate Editors

Xiao-Jun Chen , China  
Feng-Huei Lin , Taiwan  
Maria Lindén, Sweden

## Academic Editors

Cherif Adnen, Tunisia  
Saverio Affatato , Italy  
Óscar Belmonte Fernández, Spain  
Sweta Bhattacharya , India  
Prabadevi Boopathy , India  
Weiwei Cai, USA  
Gin-Shin Chen , Taiwan  
Hongwei Chen, USA  
Daniel H.K. Chow, Hong Kong  
Gianluca Ciardelli , Italy  
Olawande Daramola, South Africa  
Elena De Momi, Italy  
Costantino Del Gaudio , Italy  
Ayush Dogra , India  
Luobing Dong, China  
Daniel Espino , United Kingdom  
Sadiq Fareed , China  
Mostafa Fatemi, USA  
Jesus Favela , Mexico  
Jesus Fontecha , Spain  
Agostino Forestiero , Italy  
Jean-Luc Gennisson, France  
Badicu Georgian , Romania  
Mehdi Gheisari , China  
Luca Giancardo , USA  
Antonio Gloria , Italy  
Kheng Lim Goh , Singapore  
Carlos Gómez , Spain  
Philippe Gorce, France  
Vincenzo Guarino , Italy  
Muhammet Gul, Turkey  
Valentina Hartwig , Italy  
David Hewson , United Kingdom  
Yan Chai Hum, Malaysia  
Ernesto Iadanza , Italy  
Cosimo Ieracitano, Italy





Giovanni Improta , Italy  
Norio Iriguchi , Japan  
Mihajlo Jakovljevic , Japan  
Rutvij Jhaveri, India  
Yizhang Jiang , China  
Zhongwei Jiang , Japan  
Rajesh Kaluri , India  
Venkatachalam Kandasamy , Czech Republic  
Pushpendu Kar , India  
Rashed Karim , United Kingdom  
Pasi A. Karjalainen , Finland  
John S. Katsanis, Greece  
Smith Khare , United Kingdom  
Terry K.K. Koo , USA  
Srinivas Koppu, India  
Jui-Yang Lai , Taiwan  
Kuruva Lakshmanna , India  
Xiang Li, USA  
Lun-De Liao, Singapore  
Qiu-Hua Lin , China  
Aiping Liu , China  
Zufu Lu , Australia  
Basem M. ElHalawany , Egypt  
Praveen Kumar Reddy Maddikunta , India  
Ilias Maglogiannis, Greece  
Saverio Maietta , Italy  
M.Sabarimalai Manikandan, India  
Mehran Moazen , United Kingdom  
Senthilkumar Mohan, India  
Sanjay Mohapatra, India  
Rafael Morales , Spain  
Mehrbakhsh Nilashi , Malaysia  
Sharnil Pandya, India  
Jialin Peng , China  
Vincenzo Positano , Italy  
Saeed Mian Qaisar , Saudi Arabia  
Alessandro Ramalli , Italy  
Alessandro Reali , Italy  
Vito Ricotta, Italy  
Jose Joaquin Rieta , Spain  
Emanuele Rizzuto , Italy



Dinesh Rokaya, Thailand  
Sébastien Roth, France  
Simo Saarakkala , Finland  
Mangal Sain , Republic of Korea  
Nadeem Sarwar, Pakistan  
Emiliano Schena , Italy  
Prof. Asadullah Shaikh, Saudi Arabia  
Jiann-Shing Shieh , Taiwan  
Tiago H. Silva , Portugal  
Sharan Srinivas , USA  
Kathiravan Srinivasan , India  
Neelakandan Subramani, India  
Le Sun, China  
Fabrizio Taffoni , Italy  
Jinshan Tang, USA  
Ioannis G. Tollis, Greece  
Ikram Ud Din, Pakistan  
Sathishkumar V E , Republic of Korea  
Cesare F. Valenti , Italy  
Qiang Wang, China  
Uche Wejinya, USA  
Yuxiang Wu , China  
Ying Yang , United Kingdom  
Elisabetta Zanetti , Italy  
Haihong Zhang, Singapore  
Ping Zhou , USA

## Contents

---

**Integration of Global and Local Features for Specular Reflection Inpainting in Colposcopic Images**  
Xiaoxia Wang , Ping Li, Yuchun Lv, Huifeng Xue, Tianxiang Xu , Yongzhao Du , and Peizhong Liu 




Research Article (11 pages), Article ID 5401308, Volume 2021 (2021)

**Bayesian Fully Convolutional Networks for Brain Image Registration**  
Kunpeng Cui, Panpan Fu, Yinghao Li , and Yusong Lin 

Research Article (12 pages), Article ID 5528160, Volume 2021 (2021)

**A Semiautomated Deep Learning Approach for Pancreas Segmentation**  
Meixiang Huang , Chongfei Huang, Jing Yuan, and Dexing Kong 


Research Article (10 pages), Article ID 3284493, Volume 2021 (2021)

**Tic Detection in Tourette Syndrome Patients Based on Unsupervised Visual Feature Learning**  
Junya Wu , Tianshu Zhou, Yufan Guo, Yu Tian, Yuting Lou, Hua Ru, Jianhua Feng , and Jingsong Li 

Research Article (10 pages), Article ID 5531186, Volume 2021 (2021)

**Fp<sup>roi</sup>-GAN with Fused Regional Features for the Synthesis of High-Quality Paired Medical Images**  
Jiale Dong , Caiwei Liu , Panpan Man , Guohua Zhao, Yaping Wu , and Yusong Lin 

Research Article (13 pages), Article ID 6678031, Volume 2021 (2021)

**MAGAN: Mask Attention Generative Adversarial Network for Liver Tumor CT Image Synthesis**  
Yang Liu, Lu Meng , and Jianping Zhong

Research Article (11 pages), Article ID 6675259, Volume 2021 (2021)

## Research Article

# Integration of Global and Local Features for Specular Reflection Inpainting in Colposcopic Images

Xiaoxia Wang <sup>1</sup>, Ping Li,<sup>2</sup> Yuchun Lv,<sup>2</sup> Huifeng Xue,<sup>3</sup> Tianxiang Xu <sup>4</sup>,  
Yongzhao Du <sup>1,4</sup> and Peizhong Liu <sup>1,4</sup>

<sup>1</sup>College of Medicine, Huaqiao University, Quanzhou, Fujian 362021, China

<sup>2</sup>Department of Gynecology and Obstetrics, The First Hospital of Quanzhou, Quanzhou, Fujian 362000, China

<sup>3</sup>Cervical Disease Diagnosis and Treatment Health Center, Fujian Provincial Maternity and Children's Hospital, Affiliated Hospital of Fujian Medical University, Fuzhou, Fujian 350001, China

<sup>4</sup>College of Engineering, Huaqiao University, Quanzhou, Fujian 362021, China

Correspondence should be addressed to Peizhong Liu; pzliu@hqu.edu.cn

Received 12 May 2021; Accepted 20 July 2021; Published 28 July 2021

Academic Editor: Ayush Dogra

Copyright © 2021 Xiaoxia Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Objective.** To explore an inpainting method that can balance texture details and visual observability to eliminate the specular reflection (SR) regions in the colposcopic image, thus improving the accuracy of clinical diagnosis for cervical cancer. **Methods.** (1) To ensure smoothness, Gaussian Blur and filling methods are applied to the global image. (2) Striving to preserve the anatomical texture details of the colposcopic image as much as possible, the exemplar-based method is applied to local blocks. (3) The colposcopic images inpainted in the previous two steps are integrated, so that important information of non-SR regions is preserved based on eliminating SR regions. **Results.** In the subjective visual assessment of inpainting results, the average of 3.55 ranks first in the five comparison sets. As to the clinical test, comparing the diagnosis results of 6 physicians before and after eliminating SR regions, the average accuracy of two kinds of classifications increased by 1.44% and 2.03%, respectively. **Conclusions.** This method can effectively eliminate the SR regions in the colposcopy image and present a satisfactory visual effect. **Significance.** As a preprocessing method for computer-aided diagnosis systems, it can also improve physicians' accuracy in clinical diagnosis.

## 1. Introduction

According to World Health Organization (WHO) Global Cancer Statistics Report in 2018, cervical cancer ranked fourth in both incidence and mortality [1]. The latest statistics from the US indicate that cervical cancer remains the second leading cause of cancer deaths among women aged 20 to 39 years old [2], posing a severe threat to women's health. Clinical studies have confirmed that persistent high-risk human papillomavirus virus (HR-HPV) infection is the leading cause of the development of cervical cancer. It takes years or even decades for patients with persistent HR-HPV infection to develop from HPV infection to cervical cancer, and they also experience a long precancerous stage (CIN1, CIN2, and CIN3) [3], during which clinicians can early

detect, treat, and remove the affected tissues to prevent cervical cancer [4]. Despite the continued development of the HPV vaccine, its popularity cannot meet current needs due to its price and geographical differences. Therefore, a large-scale and standardized cervical cancer screening program for the general population is one of the most effective ways to reduce the incidence and mortality of cervical cancer.

Currently, there are three mainstream screening methods: pap cytology, colposcopy, and biopsy [5]. Among them, colposcopy has become a critical assistant tool for cervical cancer screening due to its simple operation and low cost. Colposcopy is an optical instrument that can adjust the light source to penetrate the tissue, magnify the cervical epithelium and blood vessels, and discover potential cervical



lesions and evaluate them. Therefore, when the light from the camera flash irradiates the cervical tissue during the colposcopy, some specular reflection (SR) regions often appear in colposcopic images due to the presence of physiological mucus on the surface of the cervical tissue [6]. As shown in Figure 1, in the colposcopic image, these SR regions have similar characteristics as the acetic white (AW) regions [7], which are essential tissue changes in lesion regions after the application of acetic acid. In addition, if the surface color, texture features, and saturation of the cervical tissue are weakened, the images will show high brightness and low saturation. It will result in the uneven appearance of cervical epithelial tissue and even complete loss of surface information. This phenomenon also interferes with the recognition, segmentation, and classification of cervical lesions by the computer, thereby reducing the accuracy of the cervical intelligent assistant diagnosis system. In practical applications, the preprocessing of the colposcopic image to eliminate SR regions has become an essential task for the intelligent diagnosis of cervical lesions.

## 2. Related Work

At present, many researches on the intelligent diagnosis of colposcopic images do not consider the impact of SR regions [8, 9] or just perform simple threshold processing to eliminate reflective pixels. Only a few studies on the recognition and classification of cervical lesions have considered the interference of SR regions.

Two main directions in the research on SR regions' elimination in natural images: one is the dichromatic reflection model (DRM) based on physical methods to automatically eliminate SR regions [10, 11]. It defines the color as a linear combination of object color and highlight color [12]. Another is to use polarization filters to determine SR regions [13] and then performs analysis and statistics based on the integration of multiview color and polarization information.

Since the colposcopic image contains many regions with similar colors but different intensities and textures [14], the above method cannot be fully applied to eliminating SR regions of the colposcopic image. The problem of SR regions has always been a bottleneck restricting the development of automatic extraction algorithms in the colposcopic image.

Most researchers have explored different color spaces. Van et al. [15] expressed the pixel distribution in the image as a Gaussian mixture model in the RGB color space and then distinguished SR pixels and non-SR pixels, while Praba et al. [16] performed Gaussian mixture modeling in the HIS space. Langer et al. [17] and Das et al. [18–20] performed adaptive threshold detection in RGB color space. The former mainly used the R channel, but the latter used the intersection of the three channels. Then, the smoothest interpolation and filling were performed, respectively, by the Laplacian equation. Gordon et al. [21] detected high-brightness and low-saturation regions with fixed thresholds and selected SR candidate regions to continuously refine the pixel range in the S-V space mapped from HSV color space. Zimmerman et al. [14] also adopted the same mapping space

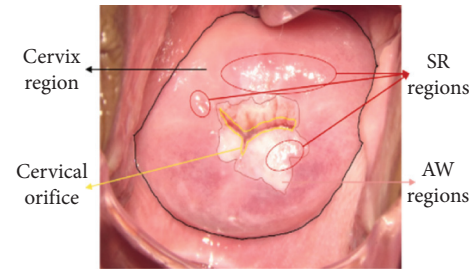


FIGURE 1: Features of cervical lesions under colposcopy.

but loosened the threshold to determine SR regions and non-SR regions more reasonably. Besides, they added the Gaussian distribution description to them and then used the iterative filling method to repair. This method requires high prior knowledge and has difficulties in actual operation. Meslouhi et al. also detected SR features in HSV color space initially [22] but later converted to planar XYZ color space, where SR regions were effectively detected by a simple automatic luminance-chromaticity comparison [23]. Here, the DRM described previously was applied. As for the inpainting, they introduced a multiresolution inpainting technique (MIT), which fully considers the different levels of details in a colposcopic image, but the method's stability needs to be verified. Kudva et al. [24] innovatively combined the SR features of HSV, RGB, and Lab color spaces to detect SR regions. Their results were stable and accurate, but the repair method was not precise.

It is worth noting that colposcopic images have many similarities with other endoscopic images, so some standard research methods can be applied to the problem of eliminating SR regions in the colposcopic images. Colonoscopy [25], thoracoscopy [26], laparoscopy [27], and more researches on eliminating SR regions in these images [28, 29] can be used for reference. In the report of Wang et al. [30], they used Arnold's method [25] in colonoscopy images and combined it with the exemplar-based method to eliminate SR regions in colposcopic images, showing better effect.

There is no ground truth without SR regions in the endoscopic image. The irregular distribution of SR pixels makes accurate manual segmentation time-consuming and labor-intensive. And manual annotation is highly dependent on experts, which increases the difficulty of quantitative analysis of the results related to SR regions elimination and limits the application of deep learning in this study, so there are relatively few studies on deep learning for eliminating SR regions in these images [31, 32].

Overall, eliminating SR regions in colposcopic images focuses on the extraction and detection of luminance and chrominance features in different color spaces, and inpainting methods are used to eliminate them. However, these methods fail to effectively balance the visual visibility and detailed preservation to better meet the subsequent processing by computers and clinical diagnosis.

This paper is a further study based on our previous studies [30]. We propose a method to eliminate SR regions based on the integration of global and local features in colposcopic images, which solves the problem that SR

regions cannot be wholly eliminated to retain the texture as much as possible. Moreover, through integrating global and local information, the overall visual observability of images is enhanced.

In the experimental part, we thoroughly investigate our method and test its performance from different angles. The results show that this method can better eliminate SR regions in colposcopic images and improve clinicians' diagnostic accuracy. The contributions of this paper are summarized as follows:

- (1) We comprehensively consider the preservation of texture details and the overall visual observability after eliminating SR regions
- (2) We propose a method to eliminate SR regions based on integrating the global and local information of the colposcopic image

### 3. Materials and Methods

Since SR regions of colposcopic images are similar to AW regions, it is not easy to directly eliminate them from chromaticity. Researchers usually detect SR regions by brightness and refine the detection by combining absolute pixels and relative pixels. However, when the detection is good, the inpainting often causes the overall image to be excessively smooth. The other way is to consider the texture details of colposcopic images, but the elimination of SR regions is incomplete, resulting in remaining highlight speckle. These two methods are not sufficiently successful in terms of visual observability. When the details of the lesion area are strictly required in clinical diagnosis, they will interfere with the physicians. With all of the above, we combine Arnold's method [25] with the exemplar-based method to eliminate SR regions of colposcopic images. At first, the original colposcopic image is smoothed by the global application of Arnold's method. Then, the colposcopic image is reconstructed after subblocking locally using the exemplar-based method to eliminate SR regions of each block finely. Finally, two images obtained in the previous steps are integrated. The overall process of our method is shown in Figure 2.

**3.1. Preprocessing.** Besides the cervical region diagnosed by physicians, there is other information in colposcopic images, including vaginal walls and other noncervical anatomy, tools (such as speculum and swabs), textual marks, and other marks superimposed on film. The elimination of SR regions is mainly aimed at the cervical region, so our SR region inpainting method is more meaningful on the preliminary preprocessing of original images.

Our data set is prenormalized. Most of the noncervical areas in original colposcopic images are trimmed, and all images are resized to a size of  $224 \times 224$  for the elimination of SR regions. In addition, to further understand whether there is a difference in images with different grades of the cervical lesion after eliminating SR regions, our data set is classified by lesion grade, and the final data sample is shown in Figure 3.

**3.2. Stage 1: Global Processing.** In the global processing, we aim at fine detection and smooth filling by learning from the method of automatic segmentation and inpainting of SR regions in colonoscopy images proposed by Arnold.

Arnold [25] proposed that, based on threshold segmentation of color images to detect SR regions, the nonlinear filtering method is used to divide the highlight pixels into two categories: significantly strong and slightly nonstrong. More accurate detection is performed gradually to avoid the influence of the background brightness. Here, according to the characteristics of high brightness and low color saturation in SR regions of colposcopic images, YUV color space transformation is performed on the image before the original algorithm to obtain the high-brightness component  $Y$  ( $C_Y$ ). Then, SR pixels are roughly and finely detected in two modules. The overall detection process of this part is shown in Figure 4. Refer to literature [25] for specific methods.

In Arnold's method [25], the gradient feature is used to limit bright non-SR regions. For colposcopic images, SR regions are usually small bright spots, while AW regions are larger white patches. Thus, to prevent certain AW regions from being recognized as SR pixels, we mainly limit the final detection regions based on the size and the brightness threshold.

For inpainting, Arnold eliminates SR regions from two levels [25]. In the first, within a certain distance of the detected edge, all the detected SR regions are replaced by the centroid color of the pixel to obtain a new modified image. Then, the modified image is filtered using the Gaussian kernel ( $\sigma = 8$ ). Finally, a robust and smooth image without SR regions is output. For the second level, a smooth weighted mask is achieved by adding a nonlinear attenuation over the contour of SR regions.

As mentioned in this literature, the larger SR regions in the image will be very blurred vision due to Gaussian Blur. For the diagnosis of colposcopic images, the requirements for texture details are strict, so if such a large blurred area appears, it will have a terrible effect in clinical practice. To solve this problem, we will introduce the exemplar-based method in the local processing, getting better texture details.

**3.3. Stage 2: Local Processing.** Local processing is mainly aimed at excessive smoothness and lack of texture details after global processing. In the previous work, we used the exemplar-based method proposed by Criminisi et al. [33] and found that this inpainting method can effectively inpaint the texture details in SR regions of colposcopic images. Therefore, we still use the exemplar-based method in this stage to balance the oversmoothing problem.

The distribution of SR regions in colposcopic images is random and uncertain. If we want to eliminate the SR region locally, the simplest solution is to block the overall image and then detect and inpaint SR regions in each block. The image is effectively processed in this way, and blocks without SR regions can be directly skipped, which reduces the time consumption to a certain extent. For the block with SR regions, the proportion of SR regions in the block is larger than the original proportion in the global image. Fortunately, the exemplar-based method works well for such a

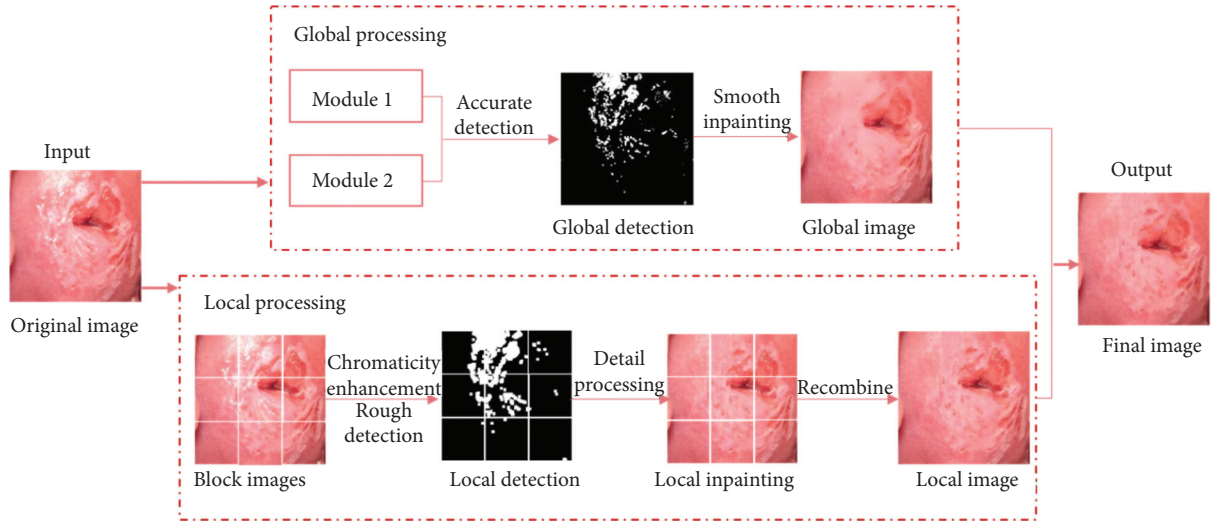


FIGURE 2: The flowchart of the proposed method.

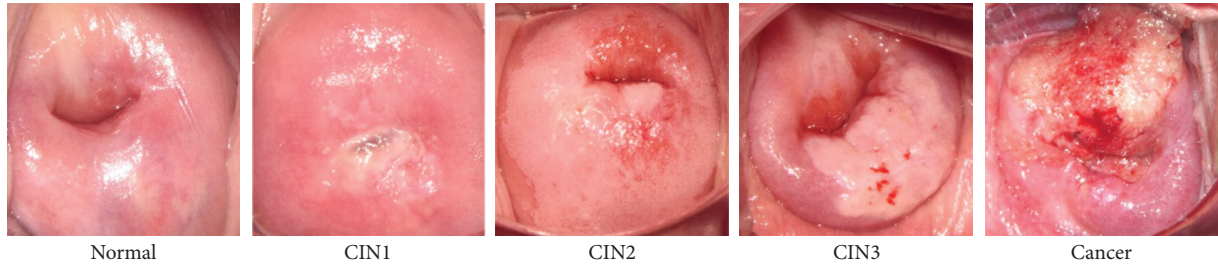


FIGURE 3: Data samples.

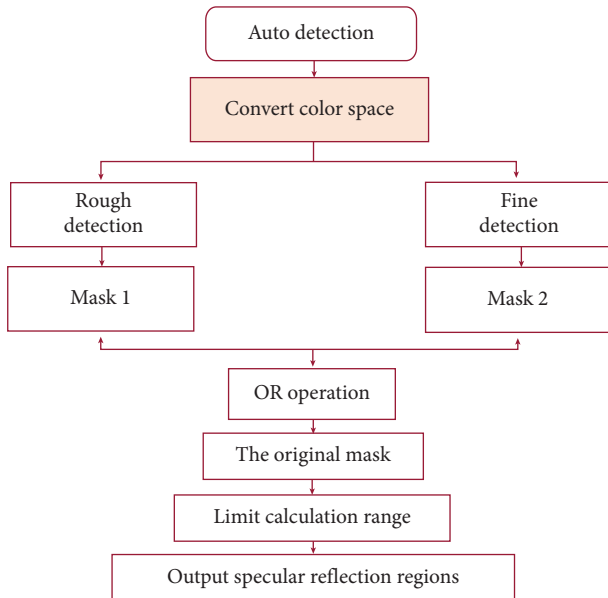


FIGURE 4: The flowchart of global SR detection.

relatively large area. In addition, when inpainting SR regions in each block, the interference from other global SR information is less, the confidence is higher, and the processing of texture detail is better. Next, we will elaborate on the detection and inpainting methods in local processing in detail.

**3.3.1. SR Regions Detection.** We require fine detection in the global processing, but our primary goal is to preserve the texture details of the image in the local processing. Therefore, we do not require precision in detecting SR regions in this part, and we require the texture details to be continuous as much as possible after inpainting. This can improve the visual observability of the colposcopic image. Based on the above considerations and the time consumption, we choose a method that uses the color characteristics of SR regions [23].

Firstly, image enhancement is performed. Since the chromaticity difference between SR regions and AW regions in the colposcopic image is small, we target the chromaticity enhancement in the HSV color space. This nonlinear filter is defined as follows:

$$\begin{pmatrix} R' \\ G' \\ C' \end{pmatrix} = (1 - S) \begin{pmatrix} R \\ G \\ C \end{pmatrix} = \frac{\min(R, G, B)}{\max(R, G, B)} \begin{pmatrix} R \\ G \\ C \end{pmatrix}. \quad (1)$$

Then, the pixel luminance ( $y$ ) and the color luminance of the entire image ( $Y_{\text{global}}$ ) are compared, and the set of pixels meeting the following conditions is defined as SR regions for local detection:

$$y > \omega Y_{\text{global}} = \omega \frac{Y}{X + Y + Z}. \quad (2)$$



Here, we give  $Y_{\text{global}}$  a coefficient with the value of  $\omega$  because we do not focus too much on the small highlight pixels in the local processing, but rather inpaint the relatively large SR regions to present an excellent visual effect. Such a process of increasing the brightness threshold also shortens the subsequent inpainting time. The selection of this coefficient will be specified in the experimental section.

**3.3.2. Inpainting.** The image inpainting algorithm based on the exemplar-based method uses the pitch as the basic unit. It uses a pixel value and a confidence value to represent the centre pixel of this pitch. After giving a priority value to it, the filling order is determined by the weight value. The best matching pitch is based on a certain matching principle to fill the texture and structure information.

The basic model of the exemplar-based method to inpaint SR regions in the colposcopic image is shown in Figure 5. Here,  $I$  represents the whole image,  $\Omega$  is the region to be inpainted,  $\partial\Omega$  is its boundary, and the pixel  $p$  is a point on the  $\partial\Omega$  boundary.  $\Psi_p$  is a rectangular neighborhood centred on the point  $p$ .  $\Phi = I - \Omega$  is the non-SR regions.

The inpainting process is as follows:

- (1) Determine the boundary of the SR region in the colposcopic image. This can provide the necessary initial information to make the inpainting gradually move from the boundary to the centre.
- (2) Calculate the priority of the target pixel  $p$ . It aims to determine the pitch to be inpainted in the SR region. The calculation formula is as follows:

$$P(p) = C(p)D(p). \quad (3)$$

$C(p)$  is the confidence item used to measure the completeness of the information in the neighborhood of pixel  $p$ . A more significant value indicates that the neighborhood of pixel  $p$  contains more available information.  $D(p)$  is the data item used to measure the location of pixel  $p$ . The greater the value is, the closer pixel  $p$  is to the decisive edge. The pitch with a higher priority value and the continuous edge will be filled in earlier to preserve the texture and structure information in SR regions.

The confidence item and data item are expressed as follows:

$$C(p) = \frac{\sum_{q \in \Psi_p \cap (I - \Omega)} C(q)}{|\Psi_p|}, \quad (4)$$

$$D(p) = \frac{|\nabla I_p^\perp \cdot n_p|}{\alpha}. \quad (5)$$

In the formula,  $|\Psi_p|$  is the area of  $\Psi_p$ ,  $\alpha$  is the normalization factor,  $n_p$  is the standard unit vector of the pixel  $p$  in the boundary direction, and  $\perp$  represents the orthogonal operator.

- (3) Select the block that best matches the SR region in the known region  $\Phi$  of the colposcopic image and fill

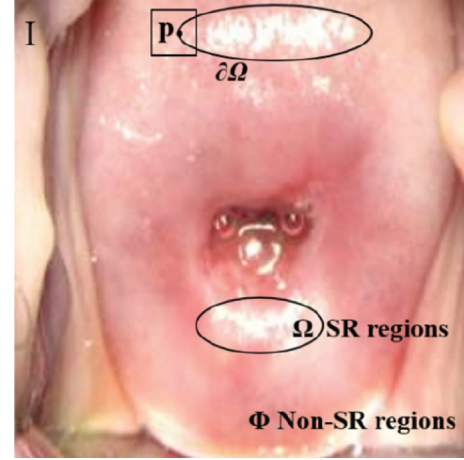


FIGURE 5: The model of exemplar-based method for SR regions inpainting.

it. According to the Sum of Squared Difference (SSD) matching principle, find the most similar pitch to copy its structure and texture information. The filling result is defined as

$$\psi_q = \arg_{\psi_q \in \Phi} \min d(\psi_p, \psi_q). \quad (6)$$

Here,  $d(\psi_p, \psi_q)$  is the sum of the squares of the differences between the corresponding pixel values in the two pitches, i.e., the relative distance. The pitch  $\psi_q$  with the smallest distance is selected as the best matching pitch of  $\psi_p$ , and the known relevant information in  $\psi_q$  is filled into the position of the SR region in the pitch  $\psi_p$ .

- (4) Update the confidence value of the corresponding pixel in the SR region. The pixel confidence values  $C(\tilde{p})$  of the filled parts of the pitch  $\psi_p$  are all replaced by the confidence value  $C(\tilde{p})$  at the centre point  $\tilde{p}$  of the pitch. The above steps are continued to be repeated after filling until all SR regions are finally eliminated.

**3.4. Stage 3: Integration.** After the first and second stages, two images focusing on global smoothing and local texture are obtained. The work of this stage is to integrate them effectively.

The colposcopic image is a color image with R, G, and B color channels. If these two images are simply added linearly in the RGB color space, the desired visual effect cannot be achieved. On the contrary, the noise information introduced in the previous inpainting process will be amplified. Therefore, we continue to analyze the colposcopic image itself and find a breakthrough in terms of luminance-chromaticity. At first, both images are converted to the HSV color space, which also contains three components, i.e., Hue (H), Saturation (S), and Value (V). Such color space is more similar to the way humans perceive colors. Then, we compare the three-component values of each pixel in the two images and reserve pixels that meet the following conditions:

$$\left\{ \begin{array}{l} H_{\text{new}} = \max(H_{\text{global}}, H_{\text{block}}) \\ S_{\text{new}} = \max(S_{\text{global}}, S_{\text{block}}) \\ V_{\text{new}} = \max(V_{\text{global}}, V_{\text{block}}) \end{array} \right\}. \quad (7)$$

The matrix of all the reserved pixels is the image we finally hope to output. Note here that, for the value of  $V$ , we also choose a more significant value instead of suppressing it, because this  $V$  here is different from SR regions that we want to eliminate in this paper. Here, we assume that SR regions have been eliminated, so the value of  $V$  more affects the brightness of the whole image. Moreover, appropriately increasing the contrast between light and shade can enhance the stereoscopy and the image's visual effect.

## 4. Results

Our experiment is divided into three sections. In Section 4.1, we set and adjust several essential parameters in eliminating SR regions. In Section 4.2, we evaluate the effect of inpainting, respectively. Finally, we invite clinicians to conduct clinical evaluation and verification in Section 4.3. All sections involving computer processing are performed on Matlab2018b. The CPU is Intel i7-8700K (3.20 GHz), and the memory is 8.00 GB.

**4.1. Parameters Setting.** In this section, we discuss the settings of several important parameters to study their impact on performance. 150 image samples (Normal, CIN1, CIN2, CIN3, Cancer, 30 in each kind) from Fujian Provincial Maternity and Children's Hospital are selected for verification. The patient information in images is processed for concealment.

**4.1.1. Block Numbers.** An innovation of this paper is the integration of global and local information, so the regulation of local scope has become a fundamental research problem of this method. In this paper, we mainly deal with local information in the form of blocks, so the determination of the local scope is controlled by the number of blocks. Under the condition of ensuring that other parameters are the same, we conduct five sets of experiments with the number of blocks of 1, 4, 9, 16, and 25 on the colposcopic image and compare the effects of detection and inpainting in turn.

As shown in Figure 6, the first column is the original image and its SR regions mask. The following columns are SR regions detected in combination with different block numbers and corresponding inpainted images. In order to enhance the readability, we have also labelled the details in the figure.

**4.1.2. The Coefficient of Local Detection.** In the local processing, we choose a simple brightness threshold of the color image. The threshold value directly affects the positioning of SR regions during the local processing and then affects their subsequent inpainting.

As mentioned above, our local detection is controlled by the brightness  $y$ . Therefore,  $y$  is treated with four different coefficients of 1.0, 1.1, 1.2, and 1.3, respectively, when the other parameters are the same. The corresponding results of detection are shown in Figure 7. Since the direct effect of this coefficient on inpainting is not obvious, here is not the comparison of inpainting. Similarly, the individual details are marked in the figure.

**4.2. Inpainting Evaluation.** To evaluate the inpainting effect, we first show eliminating results of different lesion grades in our colposcopic image dataset and then extract some literature Atlas to compare several processing methods' subjective visual evaluation grades.

**4.2.1. The Preliminary Results.** This section shows some experimental results to intuitively reflect the processing effect of our method in different grades of colposcopic images in this paper. As shown in Figure 8, images of different categories are ranked in row order, namely, Normal, CIN1, CIN2, CIN3, and cancer from top to bottom, with an example for each category. From left to right, the sequence from left to right is the resulting image of the original image, Arnold's method, Criminisi's method, the simple combination of the former two methods (Global A + C), and our proposed method integrated global and local information.

**4.2.2. Subjective Visual Evaluation.** In clinical practice, colposcopic images have no ground truth without SR regions in the real sense, so an objective quantitative evaluation cannot be effectively carried out. Thus, we construct an independent user study to provide honest feedback and quantify subjective evaluation.

Firstly, from 16 pieces of literature [17, 18, 21–23, 25, 26, 28, 31, 32, 34–39] concerning SR elimination from endoscopic images, we extracted 50 images accompanied by their corresponding result images as a new dataset, called Ref\_set. Due to the limited research specializing in SR regions elimination of colposcopic images, we extended the images to a broader range of endoscopic images, including some colonoscopic and laparoscopic images. Then, four comparison sets were generated from Ref\_set using four methods, including Arnold's method, Criminisi's method, the simple combination of the former two methods, and our proposed method integrated global and local information. Moreover, the result images in those pieces of literature were taken as another comparison set (Ref results), so a total of five comparison sets were used for subjective visual evaluation. Next, we invited 10 testers with basic knowledge of medical anatomy (Group 1) and 10 testers with experience in computer image processing (Group 2). They were required to perform a subjective visual evaluation for each image independently according to their own needs and feelings. The only instruction we gave during the evaluation is as follows: based on SR inpainting, comprehensive visual perception should be used to make a

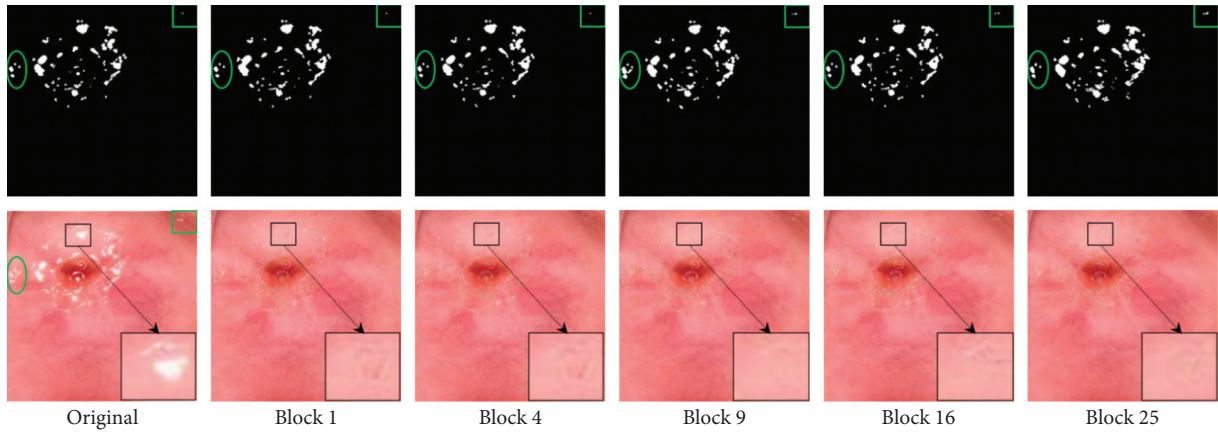


FIGURE 6: Comparison of the number of blocks.

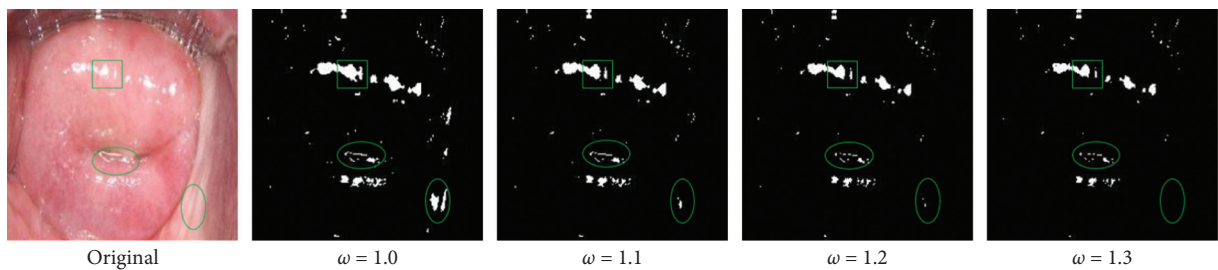


FIGURE 7: Comparison of local detection threshold coefficients.

quality rating (1 represents the worst quality, 5 represents the best quality). Finally, we make a statistical analysis of the evaluation results.

The average scores of the 1000 trials are shown in Table 1, and the specific results of the two kinds of participants are separately analyzed. In Figure 9, we also offer the corresponding scatter plots of the five processing methods and the average score of each participant.

**4.3. Clinical Test.** The ultimate goal of medical image processing is to help clinicians improve diagnosis efficiency, so we also perform clinical verification on our method in this paper.

Firstly, we select 300 colposcopic images, including 65 normal or inflammatory samples, 65 CIN1 samples, 60 CIN2 samples, 60 CIN3 samples, and 50 cervical cancer samples. Next, SR regions are eliminated by the proposed method. Then, we randomly mix 300 original images and 300 inpainting images with no SR regions to form an evaluation dataset containing 600 images. Finally, we invite 6 colposcopy physicians to make the independent diagnosis.

In the analysis results, we reparate the original images from the inpainting images and calculate the accuracy of each physician's diagnosis of colposcopic images before and after eliminating SR regions. In addition, based on the practical clinical significance, we conduct two kinds of classifications: two categories (Normal & Lesion) and four categories (Normal & LSIL & HSIL & Cancer). The statistical results are shown in Table 2. It can be seen that, after

eliminating SR regions of colposcopic images, the overall accuracy of physicians is improved to a certain extent. The accuracy of the highest one is increased by 5%. The results of the two categories mainly show positive effects, while the negative impact is slightly increased in the more refined four categories.

## 5. Discussion

This paper proposes a method to eliminate SR regions of colposcopic images and conduct various experimental tests on its performance in many aspects.

In the parameters section, we explore the number of blocks and local detection coefficients. In Figure 6, we focus on the comparison of several details of the labels. When the number of blocks is 9 and 16, the overall image restoration effect is better, but the restoration of 9 blocks takes less time. Therefore, for our dataset, in combination with the time and final effect, we believe that the performance is the best when the number of blocks is set to 9. As shown in Figure 7, with the continuous increase of the coefficient, the detected region decreases, and some SR regions with scattered distribution are lost. However, the refinement of the details is more in line with the actual SR regions positioning in the image, such as the region circled by the ellipse in the lower right corner. To balance the above two points, we apply a local detection coefficient of 1.2 to our dataset.

The overall effect of eliminating SR regions in colposcopic images is verified by subjective visual evaluation from the computer and medical perspectives to make up for the



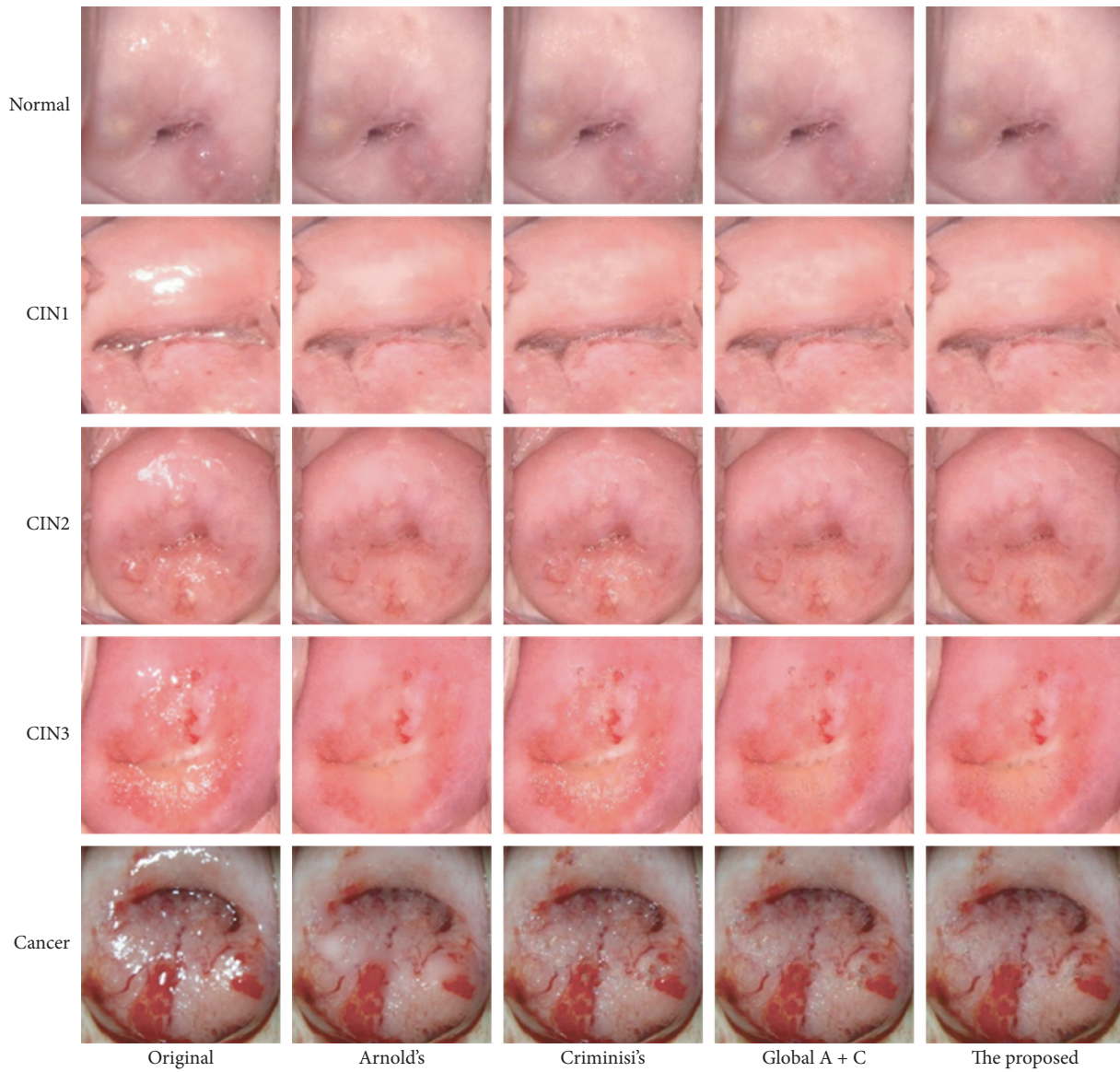


FIGURE 8: Experimental results of different grades of lesions.

TABLE 1: Subjective visual evaluation score.

	Arnold's	Criminisi's	Ref results	Global A + C	The proposed
Group 1	3.28	3.43	<b>3.63</b>	3.56	<b>3.63</b>
Group 2	2.47	2.81	3.41	3.45	<b>3.48</b>
Average	2.88	3.12	3.52	3.51	<b>3.55</b>

Bold values represent the best data in each group.

defect of comparative evaluation of no ground truth as much as possible. The evaluation results in Table 1 show that our method of integrating global and local information for eliminating SR regions in colposcopic images has the best performance from the computer processing perspective. And the effect is the same as that in the published literature from a medical standpoint. Overall, the proposed method has sound visual effects. In Figure 9, the red dots representing the proposed method are kept at the top of each

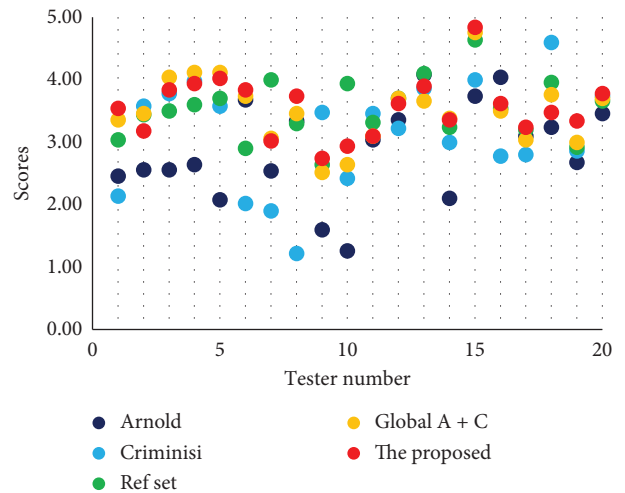


FIGURE 9: Scatter plot of subjective visual evaluation.

TABLE 2: Accuracy of clinical evaluation results (%).

Physician	Normal & lesion			Normal & LSIL & HSIL & cancer		
	Original (%)	No SR (%)	Influence (%)	Original (%)	No SR (%)	Influence (%)
1	76.67	79.00	2.33	48.33	51.67	3.34
2	69.67	71.00	1.33	48.00	50.00	2.00
3	60.67	61.67	1.00	47.00	44.67	-2.33
4	67.00	66.33	-0.67	46.33	45.67	-0.66
5	68.00	72.33	4.33	40.67	45.67	<b>5.00</b>
6	47.00	51.00	4.00	28.33	29.67	1.34
Average	64.83	66.89	2.06	43.11	44.56	1.44

Bold values represent the best data in the evaluation.

group. In a sense, this small-scale experiment makes up for the lack of ground truth to make the effective quantitative evaluation in eliminating SR regions in the colposcopic image and provides additional support for our method of integrating global and local information to eliminate SR regions.

As for clinical testing, in response to the increase in the negative impact of the four categories, we further conduct statistics on each of the four categories' accuracy and observe that the negative effect is concentrated in the Cancer category. We have feedback communication with the physicians about this result and learned that, to avoid overdiagnosis in clinical diagnosis, the physicians would maintain a conservative attitude toward a diagnosis in most cases. Ultimately, the pathological biopsy result is taken as the gold standard. Physicians often tend to make a conservative diagnosis for the images after eliminating SR regions due to insufficient brightness. This feedback also arouses our thinking. Besides the improvement of the method, the clinical application of SR elimination from colposcopic images should focus on the needs of physicians. In the process of colposcopy, when physicians find that the SR regions in the collected image interfere with the lesion diagnosis, they can select the operation of SR elimination in real-time and can also perform the comparison before and after eliminating SR regions, thus increasing the accuracy of the clinician's diagnosis.

In addition, some endoscopic datasets are involved in the detection and inpainting evaluations in this paper. The relevant evaluation results are good, proving that the method has a long-lasting effect not only for the targeted restoration of SR regions in colposcopic images, but also for other similar endoscopic images.

## 6. Conclusions

We introduce a method of eliminating SR regions by integrating the global and local information of colposcopic images in this paper. Our method preserves and constructs the texture and structure information in SR regions as much as possible, thus increasing the visual observability of the image. Many results have been achieved in computer and clinical tests. In contrast to experiments on eliminating SR regions in endoscopic images, our method still has good

performance, so it has potential value for similar visual tasks of this kind of image.

## Data Availability

The processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

The authors gratefully acknowledge the kind cooperation of Fujian Provincial Maternity and Children's Hospital during the process of cervical images acquisition. This work was supported by Quanzhou Scientific and Technological Planning Projects (grant no. 2019C028R) and in part by the grants from Fujian Provincial Science and Technology Major Project (grant no. 2020HZ02014).

## References

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020.
- [3] M. Safaeian, D. Solomon, and P. E. Castle, "Cervical cancer prevention-cervical screening: science in evolution," *Obstetrics & Gynecology Clinics of North America*, vol. 34, no. 4, pp. 739–760, 2007.
- [4] W. Kang, X. Qi, N. J. Tresser, M. Kareta, J. L. Belinson, and A. M. Rollins, "Diagnostic efficacy of computer extracted image features in optical coherence tomography of the precancerous cervix," *Medical Physics*, vol. 38, pp. 107–113, 2011.
- [5] P. E. Gravitt, P. Paul, H. A. Katki et al., "Effectiveness of VIA, Pap, and HPV DNA Testing in a cervical cancer screening program in a peri-urban community in Andhra Pradesh, India," *PLoS One*, vol. 5, no. 10, Article ID e13711, 2010.
- [6] A. G. Waxman, C. Conageski, M. I. Silver et al., "ASCCP colposcopy standards: how do we perform colposcopy?"

- implications for establishing standards,” *Journal of Lower Genital Tract Disease*, vol. 21, no. 4, pp. 235–241, 2017.
- [7] K. Fernandes, J. S. Cardoso, and J. Fernandes, “Automated methods for the decision support of cervical cancer screening using digital colposcopies,” *IEEE Access*, vol. 6, pp. 33910–33927, 2018.
- [8] B. Bai, Y. Du, P. Liu, P. Sun, P. Li, and Y. Lv, “Detection of cervical lesion region from colposcopic images based on feature reselection,” *Biomedical Signal Processing and Control*, vol. 57, pp. 101785–101792, 2020.
- [9] T. Zhang, Y.-m. Luo, P. Li et al., “Cervical precancerous lesions classification using pre-trained densely connected convolutional networks with colposcopy images,” *Biomedical Signal Processing and Control*, vol. 55, pp. 101566–101576, 2020.
- [10] H.-L. Shen, H.-G. Zhang, S.-J. Shao, and J. H. Xin, “Chromaticity-based separation of reflection components in a single image,” *Pattern Recognition*, vol. 41, no. 8, pp. 2461–2469, 2008.
- [11] Y. H. Deng, X. Y. Kong, and Q. N. Peng, “Specular highlight suppression algorithm based on chromaticity analysis and L1-weighted regularization,” *Electro Optics and Control*, vol. 25, pp. 39–42, 2018.
- [12] J. Suo, D. An, X. Ji, H. Wang, and Q. Dai, “Fast and high quality highlight removal from a single image,” *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5441–5454, 2016.
- [13] Q. Yang, J. Tang, and N. Ahuja, “Efficient and robust specular highlight removal,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1304–1311, 2015.
- [14] G. Zimmerman and H. Greenspan, “Automatic detection of specular reflections in uterine cervix images,” in *Proceedings of SPIE*, J. M. Reinhardt and J. P. W. Pluim, Eds., SPIE, San Diego, CA, USA, pp. 2037–2045, 2006.
- [15] V. V. Raad, “Frequency space analysis of cervical images using short time fourier transform,” in *Proceedings of the IASTED International Conference of Biomedical Engineering*, M. H. Hamza, Ed., Acta, Salzburg, Austria, pp. 77–81, 2003.
- [16] P. S. R. Praba and H. Ranganathan, “Wavelet transform based automatic lesion detection in cervix images using active contour,” *Journal of Computer Science*, vol. 9, no. 1, pp. 30–36, 2013.
- [17] H. Lange, “Automatic glare removal in reflectance imagery of the uterine cervix,” *Medical Imaging 2005: Image Processing*, SPIE, San Diego, CA, USA, pp. 2183–2192, 2005.
- [18] A. Das, A. Kar, and D. Bhattacharyya, “Elimination of specular reflection and identification of ROI: the first step in automated detection of cervical cancer using digital colposcopy,” in *Proceedings of the 2011 IEEE International Conference on Imaging Systems and Techniques*, pp. 237–241, IEEE, Batu Ferringhi, Malaysia, May 2011.
- [19] A. Das, A. Kar, and D. Bhattacharyya, “Preprocessing for automating early detection of cervical cancer,” in *Proceedings of the 2011 15th International Conference on Information Visualisation*, pp. 597–600, IEEE, London, UK, July 2011.
- [20] A. Das and A. Choudhury, “A novel humanitarian technology for early detection of cervical Neoplasia: ROI extraction and SR detection,” in *Proceedings of the 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 457–460, IEEE, Dhaka, Bangladesh, December 2017.
- [21] S. Gordon, G. Zimmerman, R. Long, S. Antani, J. Jeronimo, and H. Greenspan, “Content analysis of uterine cervix images: initial steps toward content-based indexing and retrieval of cervigrams,” in *Proceedings of SPIE*, J. M. Reinhardt and J. P. W. Pluim, Eds., SPIE, San Diego, CA, USA, pp. 1549–1556, 2006.
- [22] O. E. Meslouhi, H. Allali, T. Gadi, Y. A. Benksddour, and M. Kardouchi, “Image registration using opponent SIFT descriptor: application to colposcopic images with specular reflections,” in *Proceedings of the 2010 Sixth International Conference on Signal-Image Technology and Internet Based Systems*, pp. 12–17, IEEE, Kuala Lumpur, Malaysia, December 2010.
- [23] O. E. Meslouhi, M. Kardouchi, H. Allali, T. Gadi, and Y. A. Benkaddour, “Automatic detection and inpainting of specular reflections for colposcopic images,” *Central European Journal of Computer Science*, vol. 1, pp. 341–354, 2011.
- [24] V. Kudva, K. Prasad, and S. Guruvare, “Detection of specular reflection and segmentation of cervix region in uterine cervix images for cervical cancer screening,” *IRBM*, vol. 38, no. 5, pp. 281–291, 2017.
- [25] M. Arnold, A. Ghosh, S. Ameling, and G. Lacey, “Automatic segmentation and inpainting of specular highlights for endoscopic imaging,” *EURASIP Journal on Image and Video Processing*, vol. 2010, Article ID 814319, 12 pages, 2010.
- [26] C.-A. Saint-Pierre, J. Boisvert, G. Grimard, and F. Cherié, “Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images,” *Machine Vision and Applications*, vol. 22, no. 1, pp. 171–180, 2011.
- [27] J. M. Marcinczak, R.-R. Grigat, and T. Zhuang, “Closed contour specular reflection segmentation in laparoscopic images,” *International Journal of Biomedical Imaging*, vol. 2013, Article ID 593183, 6 pages, 2013.
- [28] R. Li, J. Pan, Y. Si, B. Yan, Y. Hu, and H. Qin, “Specular reflections removal for endoscopic image sequences with adaptive-RPCA decomposition,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 328–340, 2020.
- [29] F. Queiroz and T. I. Ren, “Endoscopy image restoration: a study of the kernel estimation from specular highlights,” *Digital Signal Processing*, vol. 88, pp. 53–65, 2019.
- [30] X. Wang, P. Li, Y. Du, Y. Lv, and Y. Chen, “Detection and inpainting of specular reflection in colposcopic images with exemplar-based method,” in *Proceedings of the 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, pp. 90–94, IEEE, Xiamen, China, October 2019.
- [31] D.-F. Shen, J.-J. Guo, G.-S. Lin, and J.-Y. Lin, “Content-aware specular reflection suppression based on adaptive image inpainting and neural network for endoscopic images,” *Computer Methods and Programs in Biomedicine*, vol. 192, pp. 105414–105426, 2020.
- [32] A. Rodríguez-Sánchez, D. Chea, G. Azzopardi, and S. Stabinger, “A deep learning approach for detecting and correcting highlights in endoscopic images,” in *Proceedings of the 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, IEEE, Montreal, QC, Canada, December 2017.
- [33] A. Criminisi, P. Perez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.
- [34] F. . d.S. Queiroz and T. I. Ren, “Automatic segmentation of specular reflections for endoscopic images based on sparse and low-rank decomposition,” in *Proceedings of the 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 282–289, IEEE, Rio de Janeiro, Brazil, August 2014.

- [35] S. Tchoulack, J. M. P. Langlois, and F. Cheriet, "A video stream processor for real-time detection and correction of specular reflections in endoscopic images," in *Proceedings of the 2008 Joint 6th International IEEE Northeast Workshop on Circuits and Systems and TAISA Conference*, pp. 49–52, IEEE, Montreal, QC, Canada, June 2008.
- [36] D. B. Patil, M. S. Gaikwad, D. K. Singh, and T. S. Vishwanath, "Semi-automated lesion grading in cervix images with specular reflection removal," in *Proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT)*, pp. 1–5, IEEE, Coimbatore, India, August 2016.
- [37] Y. Gao, J. Yang, S. Ma et al., "Dynamic searching and classification for highlight removal on endoscopic image," *Procedia Computer Science*, vol. 107, pp. 762–767, 2017.
- [38] G. Karapetyan and H. Sarukhanyan, "Automatic detection and concealment of specular reflections for endoscopic images," in *Proceedings of the Ninth International Conference on Computer Science and Information Technologies Revised Selected Papers*, pp. 1–8, IEEE, Yerevan, Armenia, September 2013.
- [39] S. M. Alsaleh, A. I. Aviles, P. Sobrevilla, A. Casals, and J. K. Hahn, "Adaptive segmentation and mask-specific Sobolev inpainting of specular highlights for endoscopic images," in *Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1196–1199, IEEE, Orlando, FL, USA, August 2016.



## Research Article

# Bayesian Fully Convolutional Networks for Brain Image Registration

Kunpeng Cui,<sup>1,2</sup> Panpan Fu,<sup>3</sup> Yinghao Li ,<sup>3,4</sup> and Yusong Lin ,<sup>2,3,4</sup>

<sup>1</sup>School of Information Engineering, Zhengzhou University, Zhengzhou 450001, Henan, China

<sup>2</sup>Collaborative Innovation Center for Internet Healthcare, Zhengzhou University, Zhengzhou 450052, Henan, China

<sup>3</sup>School of Software, Zhengzhou University, Zhengzhou 450002, Henan, China

<sup>4</sup>Hanwei IoT Institute, Zhengzhou University, Zhengzhou 450002, Henan, China

Correspondence should be addressed to Yinghao Li; [yinghaoli@zzu.edu.cn](mailto:yinghaoli@zzu.edu.cn) and Yusong Lin; [yslin@ha.edu.cn](mailto:yslin@ha.edu.cn)

Received 28 February 2021; Revised 17 June 2021; Accepted 13 July 2021; Published 27 July 2021

Academic Editor: Jialin Peng

Copyright © 2021 Kunpeng Cui et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The purpose of medical image registration is to find geometric transformations that align two medical images so that the corresponding voxels on two images are spatially consistent. Nonrigid medical image registration is a key step in medical image processing, such as image comparison, data fusion, target recognition, and pathological change analysis. Existing registration methods only consider registration accuracy but largely neglect the uncertainty of registration results. In this work, a method based on the Bayesian fully convolutional neural network is proposed for nonrigid medical image registration. The proposed method can generate a geometric uncertainty map to calculate the uncertainty of registration results. This uncertainty can be interpreted as a confidence interval, which is essential for judging whether the source data are abnormal. Moreover, the proposed method introduces group normalization, which is conducive to the network convergence of the Bayesian neural network. Some representative learning-based image registration methods are compared with the proposed method on different image datasets. Experimental results show that the registration accuracy of the proposed method is better than that of the methods, and its antifolding performance is comparable to that of fast image registration and VoxelMorph. Furthermore, the proposed method can evaluate the uncertainty of registration results.

## 1. Introduction

Image registration is an image-processing process that aligns two or more images of the same scene captured at different times and different perspectives or by using different sensors [1, 2]. Nonrigid medical image registration is a key step in medical image processing. In clinical diagnosis, it can judge a patient's progress by aligning the brain magnetic resonance images of the patient with Alzheimer's disease at different periods [3, 4]. In tumor surgery, rapid medical image registration can aid doctors in surgical navigation [5–7]. In demography research, image registration is helpful for studying differences in the brain tissue structures of people from different countries.

With the advances in medical image registration technology, various registration methods have been developed,

such as elastic body models [8–10], viscous fluid flow models [11–13], diffusion models [14], curvature registration [15], statistical parameter mapping [16], free-form deformation with b-spline [17], discrete method [18, 19], and demons [20] for registration model construction. Many optimization algorithms have also been devised, such as gradient descent methods [21], conjugate gradient methods [22, 23], Powell's conjugate direction method [24, 25], quasi-Newton methods [26, 27], Gauss-Newton method [28, 29], and stochastic gradient descent methods [30, 31]. Similarity measurement methods, such as the sum of squared differences [32], the sum of absolute differences, cross-correlation [33], and mutual information [34], have been proposed.

However, traditional registration methods face real-time challenges. Large amounts of input data must be processed when performing nonrigid registration modeling on 3D data

with high resolution, a step that requires a long time. The optimization part usually uses an iterative algorithm, thereby further increasing the total time needed to obtain the final result [35].

With the development of deep learning, several researchers have proposed deep neural networks to learn features of unregistered images. Registration methods based on deep learning can be supervised [36, 37] or unsupervised [38–40]. Most supervised registration methods rely on anatomical labels. However, marking anatomical labels is difficult, a step that not only consumes a lot of time of experts but also sometimes hardly guarantees accuracy. In practice, supervised registration methods are restricted. In their place, some scholars have proposed unsupervised medical image registration methods.

Several unsupervised medical image registration methods have been proposed. *VoxelMorph*, a recently proposed unsupervised learning-based method for deformable medical image registration, has a better registration accuracy and a faster speed than other registration methods [41]. Some researchers combined the advantages of classical methods and learning-based methods to produce a probabilistic generative model and derive a diffeomorphic inference algorithm [42]. A registration method called fast image registration (FAIM) for 3D medical image registration has been proposed. Compared with the registration network based on U-net, FAIM has fewer trainable parameters to obtain a higher registration accuracy. In addition, FAIM has less irreversible regions because of the penalty loss for negative Jacobian determinants [43]. Some scholars recently proposed the Probab-Mul registration method, which is a feature-level probability model that can perform regularization on the hidden layers of two deep convolutional neural networks [44].

The focus of the present study is mainly on the accuracy of registration methods and barely on the uncertainty of their registration results. The uncertainty of registration results is very important in clinical applications as it can be used to judge whether the registration result is meaningful. For example, if a model is modeling normal human brain images, it will never see abnormal brain images that have brain tumors, malformations, and edema. When the uncertainty of a registration result is higher than a certain threshold, the source image can be judged as an abnormal brain image. During testing, the Bayesian neural network can obtain the uncertainty of results. Bayesian neural networks are used in autonomous driving, classification tasks, and segmentation tasks. Some researchers recently applied Bayesian neural networks to image registration. Deshpande et al. employed a Bayesian deep learning approach for deformable medical image registration. They reported that this approach has a better performance than existing state-of-the-art approaches [45]. Khawaled et al. developed a fully Bayesian framework for unsupervised deep learning-based deformable image registration. Their approach provided better estimates of the deformation field, thereby improving registration accuracy [46]. However, these aforementioned methods do not sufficiently consider and discuss the

uncertainty of registration results. Furthermore, they are suitable for 2D images only.

In this paper, a method based on Bayesian fully convolutional networks is proposed for image registration. The proposed method generates a geometric uncertainty map to measure the uncertainty of registration results. Thus, when the source image obtained is abnormal data, the model will provide a hint that the source image is problematic instead of immediately accepting the registration result of the model. Group normalization (GN) is also added in networks. GN groups channel similar features into one group. Hence, GN can make the model easier to optimize and converge to improve registration accuracy. The performance of the registration model in evaluating uncertainty is determined.

This paper is organized as follows. Section 2 introduces the principle of the proposed method. Section 3 describes the experimental setup. Section 4 discusses the experimental results. Finally, Section 5 summarizes the results of the study and considers directions for future work.

## 2. Methods

Figure 1 presents an overview of the proposed method. We used CNN to model the function  $g_\gamma(S, T) = u$ , where  $\gamma$  is the parameter of the convolutional layers,  $S$  is the source image,  $T$  is the target image, and  $u$  is the displacement field between the source image and the target image.  $S$  and  $T$  are defined over a 3D spatial domain  $\Omega \subset \mathbb{R}^3$ . For each voxel  $p \in \Omega$ ,  $u(p)$  is the displacement, where the map  $\phi = Id + u$  is formed using an identity transform and  $u$ . The network takes  $S$  and  $T$  as the input and uses a set of parameters  $\gamma$  to calculate  $\phi$ . We used a spatial transformation function to warp  $S$  to  $S \circ \phi$  and evaluate the similarity between  $S \circ \phi$  and  $T$ . During testing, given the images  $T$  and  $S$  of the test set, we obtained the registration field by evaluating  $g_\gamma(S, T)$ .

**2.1. Architecture.** In this section, the architecture of the convolutional neural network used in the proposed method is described in detail (Figure 2). During training, the moving image and the target image are stacked together as the input fed into the Bayesian fully convolutional network module (BFCNM) [43]. The first layer is inspired by Google’s inception module. The purpose of this layer is to compare and capture information on different spatial scales of later registration. Parametric rectified linear unit [47] activation is utilized at the end of each convolution block, and linear activation is employed in the last layer to generate the displacement field. Instead of inserting max-pooling layers, a kernel stride of 2 is used to reduce image size. Three “add” skip connections are present in downsampling and upsampling [43]. The “add” skip connection is conducive to the fusion of upsampling information and its corresponding downsampling information. During the upsampling phase, two Bayesian blocks are used. The Bayesian blocks are composed of a transposed convolutional layer, a convolutional layer, PReLU, a group normalization layer, and a Dropout layer. The detail of the Bayesian block is shown in Figure 2(b). In this paper, Monte Carlo Dropout (MC-

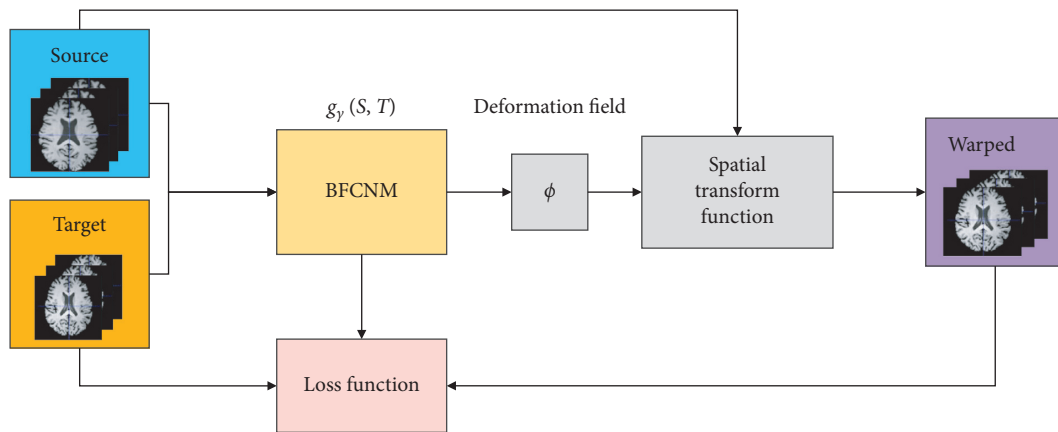


FIGURE 1: Overview of the method.

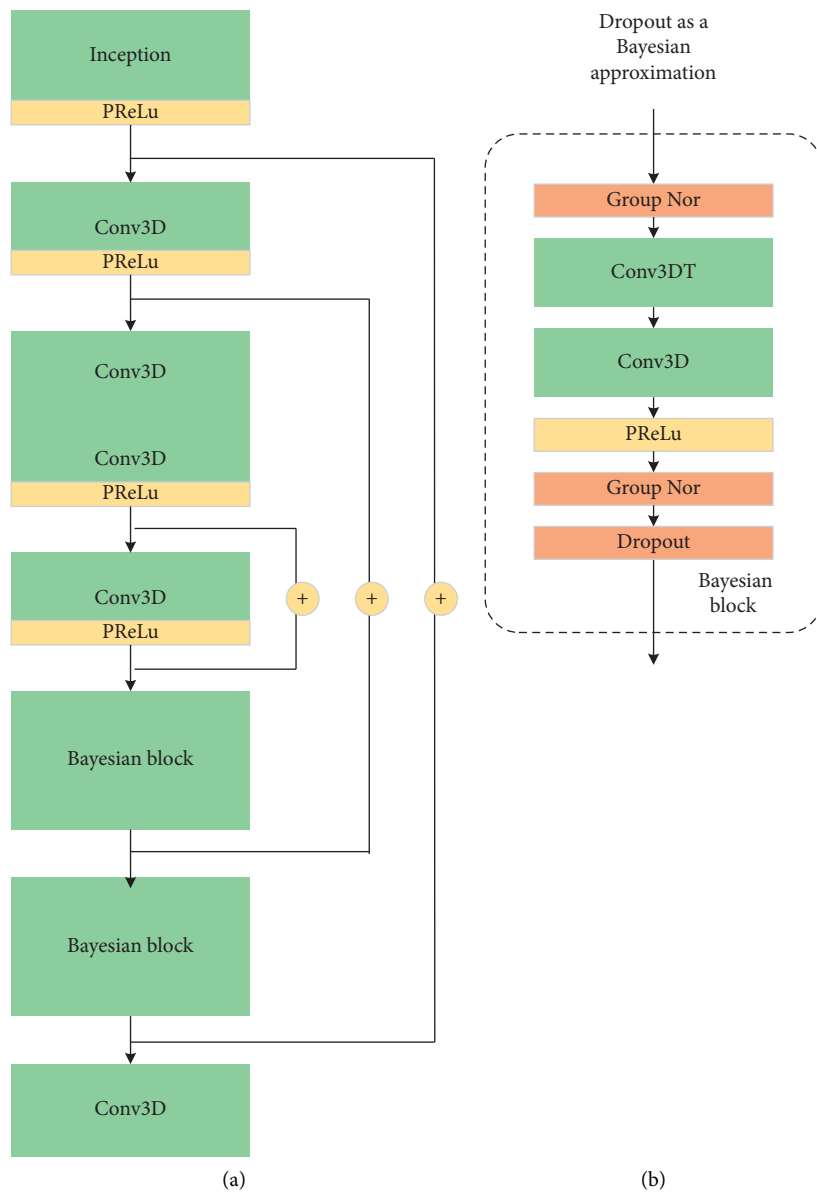


FIGURE 2: Convolution network architecture implementing BFCNM ( $g_y(S, T)$ ). (a) Highest-level view, showing sequential Conv3D and Bayesian block. (b) Details of Bayesian block.



Dropout) is introduced; it is interpreted as a Bayesian approximation of Gaussian processes.

**2.2. Spatial Transformation Function.** The spatial transformation function uses the  $\phi$  generated by BFCNM to resample  $S$  and obtain the warped image  $S \circ \phi$ . The proposed method learns the optimal parameter values by minimizing the difference between  $S \circ \phi$  and  $T$ . A differentiable operation is constructed on the basis of the spatial transformer network [41, 48] via the standard gradient-based method to calculate  $S \circ \phi$ . For each voxel  $p$ , a voxel position  $p' = p + u(p)$  is calculated in the source image. The image values are only defined in integer positions. Thus, linear interpolation is performed at eight adjacent voxels:

$$S \circ \phi(p) = \sum_{q \in Z(p')} S(q) \prod_{d \in \{x, y, z\}} (1 - |p'_d - q_d|). \quad (1)$$

where  $Z(p')$  is the voxel neighbors of  $p'$  and  $d$  is iterated in the dimension of  $\Omega$ . Errors can be backpropagated during the optimization process because gradients or subgradients can be calculated.

**2.3. Loss Function.** The total training loss is the sum of an image dissimilarity term  $L_{\text{image}}$  and the regularization terms, as shown in Table 1. The loss function [43] is defined as follows:

$$L_{\text{total}} = L_{\text{image}}(S, T) + \alpha R_1(u) + \beta R_2(u). \quad (2)$$

The main loss  $L_{\text{image}}$  with cross-correlation (CC) in this paper is for the similarity between the warped source image and the target image. The definition of CC is as follows:

$$\text{CC}(t, s \circ \phi) = \frac{(\sum_{x \in \Omega} (t(x) - \bar{t}(x))(s \circ \phi(x) - \overline{s \circ \phi}(x)))^2}{(\sum_{x \in \Omega} (t(x) - \bar{t}(x))^2)(\sum_{x \in \Omega} (s \circ \phi(x) - \overline{s \circ \phi}(x))^2)}, \quad (3)$$

where  $t(x)$  is the grey value of the target image,  $\bar{t}(x)$  is the average grey value of the target image,  $s \circ \phi(x)$  is the grey value of the warped image, and  $\overline{s \circ \phi}(x)$  is the average grey value of the warped image.

The first regularization term  $R_1$  regularizes the overall smoothness of the predicted displacements. The parameter of the regular term is  $\alpha$ , and its value is always 1. The purpose of the second regularization is to penalize transformations that have many negative Jacobian determinants. The parameter of the regular term is  $\beta$ . The transformations of all nonnegative Jacobian determinants will not be penalized. If the Jacobian determinant is negative, then the transformation result will be folded, which is physically unrealistic.

**2.4. Group Normalization.** Group normalization (GN) is a feature-normalization technique that is inserted into the architecture of deep neural networks as a trainable process. The purpose of GN is to reduce internal covariant shifts. With training iterations, the distribution of features often continuously changes. Under this condition, the parameters in the

TABLE 1: Loss and regularization functions used.

$L_{\text{image}}(S, T): 1 - \text{CC}(S \circ \phi^{-1}, T)$
Regularization : $R_1(u) = \ Du\ _2$
Regularization : $R_2(u) = 0.5( \det(D\phi^{-1})  - \det(D\phi^{-1}))$

convolutional layer must be continuously updated to adapt to the changes in distribution. GN normalizes the feature to a fixed distribution (mean value is zero, and the standard deviation is 1) and then adjusts the feature to an ideal distribution, which is learned in the training process [48].

Here,  $x$  is the feature computed by a layer, and  $i$  is an index. In the case of 3D images,  $i = (i_N, i_C, i_D, i_H, i_W)$  is a 5D vector indexing the features in  $(N, C, D, H, W)$  order, where  $N$  is the batch axis;  $C$  is the channel axis; and  $D, H,$  and  $W$  are the spatial depth, height, and width axes, respectively.

Formally, the group normalization layer must compute for mean  $\mu$  and standard deviation  $\sigma$  in a set  $S_i$ .  $S_i$  is a group and defined as follows:

$$S_i = \left\{ k | k_N = i_N, \lfloor \frac{k_C}{C/G} \rfloor = \lfloor \frac{i_C}{C/G} \rfloor \right\}, \quad (4)$$

where  $G$  is the number of groups, which is a predefined hyperparameter;  $C/G$  is the number of channels in each group;  $\lfloor \cdot \rfloor$  represents floor operation;  $\lfloor k_C / (C/G) \rfloor = \lfloor i_C / (C/G) \rfloor$  means that the indexes  $i$  and  $k$  are in the same group of channels, assuming that each group of channels is stored in sequential order along the  $C$  axis; and  $S_i$  contains all voxels along the  $(D, H, W)$  axes and along with a group of  $(C/G)$  channels.

The mean  $\mu_i$  and standard deviation  $\sigma_i$  of  $S_i$  are computed as follows:

$$\begin{aligned} \mu_i &= \frac{1}{m} \sum_{k \in S_i} x_k, \\ \sigma_i &= \sqrt{\frac{1}{m} \sum_{k \in S_i} (x_k - \mu_i)^2 + \varepsilon}, \end{aligned} \quad (5)$$

where  $\varepsilon$  is a small constant, and  $m$  is the size of set  $S_i$ . GN then performs the following computation:

$$\hat{x}_i = \frac{1}{\sigma_i} (x_i - \mu_i), \quad (6)$$

GN learns a per-channel linear transform to compensate for the possible loss of representational ability:

$$y_i = \gamma \hat{x}_i + \beta, \quad (7)$$

where  $\gamma$  and  $\beta$  are trainable scale and shift, respectively. Given the  $S_i$  in (4), the GN layer is defined by equations (5)–(7). Specifically, the voxels in the same group are normalized by the same  $\mu_i$  and  $\sigma_i$ . GN also learns the  $\gamma$  and  $\beta$  of each channel.

**2.5. Bayesian Neural Network.** In this section, the registration network based on Bayesian inference is introduced. The credibility of the results is important in solving medical

problems. Several researchers proposed a Bayesian neural network by studying the uncertainty of deep learning [49, 50]. The Bayesian neural network is a statistical model derived from the perspective of probability. The parameters in the model are initialized by an a priori distribution, and the parameters are further optimized by Bayesian inference. In the Bayesian neural network, given the data set  $D$  and weight  $W$ , the dataset  $D$  contains data  $X$  and label  $Y$ . The goal of Bayesian neural network training is to optimize the parameters, that is, to seek the posterior distribution of weight  $W$ . According to Bayesian criterion, the posterior distribution of weight  $W$  is written as

$$\begin{aligned} p(W|D) &= \frac{p(D|W)p(W)}{p(D)} \\ &= \frac{p(y|x, W)p(W)}{p(y|x)}, \end{aligned} \quad (8)$$

where  $x$  and  $y$  are the data in the training set and the corresponding label, respectively, and  $p(W)$  is the initial value of the parameter (i.e., the prior distribution). The posterior distribution of the labels predicted by the Bayesian neural network can then be calculated as follows [51]:

$$\begin{aligned} p(y^*|x^*, D) &= \mathbb{E}_{p(W|D)} [p(y^*|x^*, W)] \\ &= \int p(y^*|x^*, W)p(W|D)dW. \end{aligned} \quad (9)$$

In equation (9), the weight parameter  $W$  in the network is used to predict the unknown distribution of label  $y^*$ . In the Bayesian neural network model, the solution of the posterior distribution  $p(W|D)$  of the parameters is the key to the entire model. However, this solution is computationally intractable for neural networks of any size. Therefore, many researchers use approximate methods to obtain the solution [52, 53].

A common approach is to use variational inference to approximate the posterior distribution of the weights. This method introduces the variational distribution  $q_\theta(W)$  of weight  $w$ , which is parameterized on  $\theta$ . The approximate posterior distribution  $q_\theta(W)$  is obtained by minimizing the Kullback-Leibler (KL) divergence between  $q_\theta(W)$  and the true posterior distribution  $p(W|D)$ .

$$\text{KL}(q_\theta(W) \| p(W|D)). \quad (10)$$

Minimizing KL divergence is equivalent to minimizing the Negative Evidence Lower Bound (NELBO):

$$\begin{aligned} \text{NELBO} &= \mathbb{E}_{q_\theta} [-\log p(Y|X, W)] \\ &\quad + \text{KL}(q_\theta(W) \| p(W)) \\ &= - \int q_\theta(W) \log p(Y|X, W) dW \\ &\quad + \text{KL}(q_\theta(W) \| p(W)), \end{aligned} \quad (11)$$

with respect to the variational parameter  $\theta$ . The first term (commonly referred to as the expected log-likelihood) encourages  $q_\theta(W)$  to place its mass on the configurations

of the latent variable that explains the observed data. However, the second term (referred to as prior KL) encourages  $q_\theta(W)$  to be similar to the prior distribution  $p(W)$ , preventing the model from overfitting. The goal is to develop an explicit and accurate approximation for the expectation.

Our approach uses Bernoulli approximating variational inference and Monte Carlo sampling [54]. In practice, Dropout is used for Bayesian neural network approximation.

When Dropout [55] is applied to the output of a layer, the output can be written as

$$a_i^{\text{DO}} = \sigma(z_i \odot (W_i v)). \quad (12)$$

where, for a single  $K_{i-1}$  dimensional input  $v$ , the  $i^{\text{th}}$  layer of a neural network with  $K_i$  units would output a  $K_i$  dimensional activation vector;  $w_i$  is the  $K_i \times K_{i-1}$  weight matrix;  $\sigma(\cdot)$  is the nonlinear activation function;  $\odot$  signifies the Hadamard product;  $z_i$  is a  $K_i$  dimensional binary vector with its elements drawn independently from  $z_i^{(k)} \sim \text{Bernoulli}(p_i)$   $k=1, \dots, K_i$ ; and  $p_i$  is the probability of keeping the output activation.

The solution of the posterior distribution  $p(W|D)$  of the parameters is further improved after introducing the Bernoulli distribution into the weight parameters of our model. The Monte Carlo sampling method is used to estimate the first item in (11):

$$\begin{aligned} \mathbb{E}_{q_\theta} \log p(Y|X, W) &= \sum_{n=1}^N \int q_\theta(W) \log p(y_n|x_n, W) \\ &= \frac{1}{N} \sum_{n=1}^N \log p(y_n|x_n, \hat{W}_n), \end{aligned} \quad (13)$$

where  $\hat{W}_n$  is not the maximum posterior estimation but the random variable realizations from the Bernoulli distribution; and  $\hat{W}_n \sim q_\theta(W)$ , which is the same as applying Dropout to the weights of the network. For the second item in equation (11) (i.e., KL term), the approximate solution is given in the literature [56]. The KL term has been shown to be equivalent to  $\sum_{i=1}^L \|W_i\|_2^2$ . Thus, equation (11) can be rewritten as

$$\text{NELBO} = -\frac{1}{N} \sum_{n=1}^N \log p(y_n|x_n, \hat{W}_n) + \sum_{i=1}^L \|W_i\|_2^2. \quad (14)$$

Equation (14) is the unbiased estimation of equation (11). Interestingly, it is the same as the loss function used in standard neural networks with  $L_2$  weight regularization, and Dropout is applied to all weights of the network. Therefore, training such a neural network with stochastic gradient descent has the same effect as minimizing the KL term in (10). This scheme is similar to a Bayesian neural network and can generate a set of parameters that can best explain the observed data while preventing overfitting.

Predictions in this model follow (9) replacing the posterior  $p(W|D)$  with the approximate posterior  $q_\theta(W)$ . The integral can be approximated with Monte Carlo integration [51, 54]:

$$\begin{aligned}
p(y^*|x^*, \mathcal{D}) &\approx \int p(y^*|x^*, W)q_\theta(W)dW \\
&\approx \frac{1}{T} \sum_{t=1}^T \log p(y_n|x_n, \hat{W}_t) \\
&\approx p_{MC}(y^*|x^*),
\end{aligned} \tag{15}$$

where  $\hat{W}_t \sim q_\theta(W)$ , which means that, at test time, the Dropout layers are kept active to keep the Bernoulli distribution over the network weights. This integration is referred to as the Monte Carlo Dropout.

The Monte Carlo Dropout reflects the need to conduct multiple forward propagation processes on the same input. In this manner, the output of “different network structures” can be obtained under the action of Dropout during testing. The prediction results and the uncertainty of the model can be obtained by calculating the average and statistical variance of these outputs. The advantage of Bayesian deep learning is that Monte Carlo Dropout can give a prediction value and the confidence of the predicted value.

**2.6. Measuring Model Uncertainty.** Uncertainties in a network are a measure of how certain the model is with its prediction. In general, Bayesian modeling has two types of uncertainty. Model uncertainty, also known as Epistemic uncertainty, measures what the model does not know owing to the lack of training data. This uncertainty can be reduced with more data. During testing, model uncertainty can measure whether the testing data exists in the distribution of the training data. Aleatoric uncertainty measures the noise inherent in the observation data and cannot be reduced by collecting more data [51].

By computing the result of stochastic forward passes of the Bayesian neural network, the model’s confidence of its output can be estimated. In this paper, the mean  $\mu$  and the standard deviation  $\sigma$  of all displacements produced by Monte Carlo sampling are calculated. The mean  $\mu$  is used in the registration image, whereas the standard deviation  $\sigma$  provides an estimate of the uncertainty of registration results. The mean  $\mu$  of the displacement fields is calculated as follows:

$$\mu = \frac{1}{M} \sum_{i=1}^M y_i, \tag{16}$$

where  $M$  represents the number of Monte Carlo sampling ( $M = 48$  in this paper).  $y_i$  represents the displacement field sampled by  $i$ th. After calculating the mean value of the displacement fields, the standard variance of the displacement fields can be calculated as follows:

$$\sigma = \sqrt{\frac{1}{M-1} \sum_{i=1}^M (y_i - \mu)^2}, \tag{17}$$

where  $\sigma$  can be expressed as the uncertainty of registration results.

### 3. Experiment

**3.1. Experimental Setup.** The dataset we adopted herein was created by Arno et al. who based it on a collection of 101 T1-weighted MRIs from healthy subjects [57]. In this paper, we used brain images from the four subsets of Mindboggle101, namely, NKI-RS-22, NKI-TRT-20, MMRR-21, and OASIS-TRT-20, for a total of 83 images. These images are already warped to MNI152 space. Each image had a dimension of  $182 \times 218 \times 182$ , each of which we truncated to  $144 \times 180 \times 144$ . In the preprocessing stage, we utilized the FMRIB Software Library (FSL) to perform affine registration on NKI-RS-22, NKI-TRT-20, MMRR-21, and OASIS-TRT-20. We initially normalized the voxel intensity of each brain image and then normalized voxel intensity to 0–255. Finally, we performed a registration test on the five main anatomical regions of the cerebral cortex.

#### 3.2. Evaluation Metrics

**3.2.1. Dice Scores.** If the registration field  $\phi$  represents an accurate correspondence, then the corresponding anatomical regions in  $T$  and  $S \circ \phi$  should overlap well. Therefore, we evaluated registration accuracy by using the Dice score. The Dice score is defined as follows [43]:

$$\text{DICE} = \frac{2 * |X \cap Y|}{|X| + |Y|}. \tag{18}$$

**3.2.2. Regularized Penalty Folding.** We also evaluated the regularity of deformation fields. Specifically, the Jacobian matrix captures the local properties of  $\phi$  around voxel  $p$ . We counted all nonbackground voxels where the Jacobian determinant  $\det(\nabla\phi) < 0$  is negative [43]:

$$N := \sum \delta(\det(D\phi) < 0), \tag{19}$$

where  $\delta(\cdot)$  indicates that if it is true, then the return value is 1.

**3.2.3. Uncertainty Evaluation Metrics.** We adopted the method proposed in the literature to evaluate uncertainty performance [51]. We used metrics that incorporate the ground truth label, model prediction, and uncertainty value to evaluate the performance of such models in estimating uncertainty. Figure 3 shows the required processing steps to prepare these quantities for our metrics in a registration example. We computed the map of correct and incorrect values by matching the ground truth labels and the model predictions. We converted the uncertainty map into a map of certain and uncertain predictions by setting the uncertainty threshold  $T$ , which varies between the minimum and the maximum uncertainty values in the entire test set. The following indicators can reflect the characteristics of a good uncertainty estimator.

Negative predictive value (NPV): in the output of certain results by the model, NPV is the percentage of voxels that is

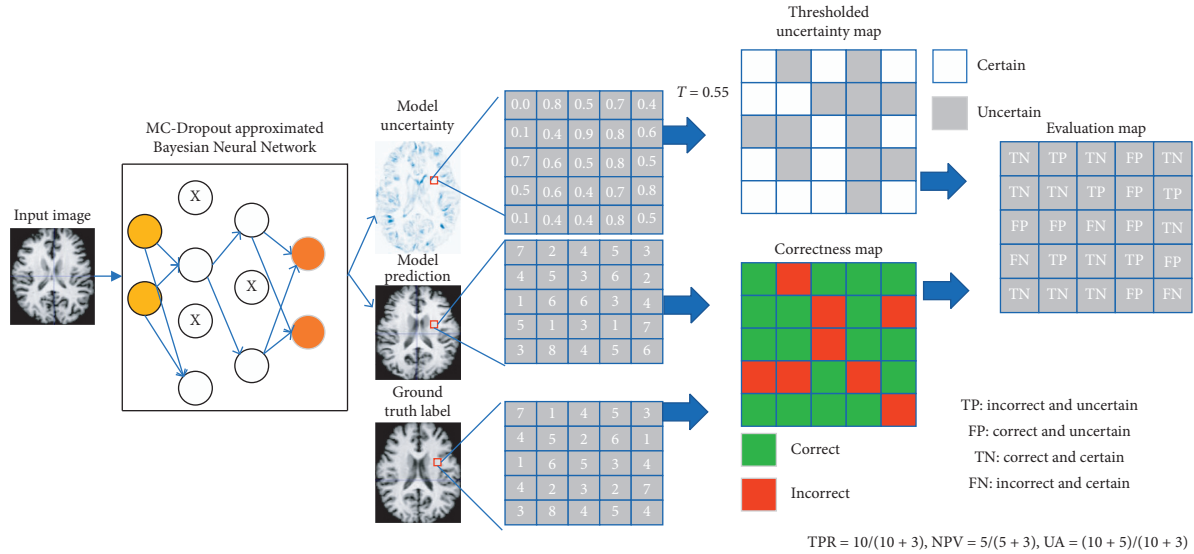


FIGURE 3: Overview of the metrics for the evaluation of the uncertainty quality in a registration example.

correctly predicted and can be written as a conditional probability:

$$NPV = \frac{P(\text{correct, certain})}{P(\text{certain})} = \frac{TN}{TN + FN}. \quad (20)$$

True positive rate (TPR): if a model is making an incorrect prediction, then the proportion of uncertain voxels is called TPR. TPR can be written as a conditional probability:

$$TPR = \frac{P(\text{uncertain, incorrect})}{P(\text{incorrect})} = \frac{TP}{TP + FN}. \quad (21)$$

Uncertainty accuracy (UA): UA is the overall accuracy of uncertainty estimation and can be measured as the ratio of the desired cases explained above (TP and TN) over all possible cases:

$$UA = \frac{P(\text{correct, certain}) + P(\text{uncertain, incorrect})}{P(\text{correct}) + P(\text{incorrect})} \quad (22)$$

$$= \frac{TP + TN}{TP + TN + FP + FN}.$$

Clearly, in all the metrics proposed above, higher values indicate that the model performs better. The values of these metrics depend on the uncertainty threshold.

**3.3. Baseline Methods.** In the comparative study, we used FSL, a comprehensive library of analytical tools for fMRI, MRI, and diffusion tensor imaging brain imaging data, as the baseline to perform an affine registration experiment with 12 degrees of freedom on the test set. We used the second baseline symmetric normalization (SyN) with mutual information as a similarity measure in the publicly available Advanced Normalization Tools (ANTs) software package [58]. We also tested the recently developed CNN-based methods, namely, VoxelMorph [41], FAIM [43], and Probab-Mul [44], and compared their performance with that of

the proposed method. The hyperparameters of the CNN-based methods were consistent. Finally, we adopted various methods for ablation study. The method that only adds GN was denoted as Our-GN, and the method that only adds Dropout was labeled as Our-DO. These two methods were consistent with our method in terms of hyperparameter settings.

**3.4. Implementation.** We divided the data set into training and test image sets. The training set consisted of all ordered brain image pairs from the union of the NKI-RS-22, NKI-TRT-20, and MMRR-21 subsets, which comprised 3906 pairs in total. The test set consisted of all ordered pairs from the OASIS-TRT-20 subset with 380 pairs in total. We trained FAIM, VoxelMorph, Probab-Mul, and our method on all pairs of images from the training set and then examined their predicted deformations by using the pairs of images from the testing set.

We implemented our method using Keras [59] with a Tensorflow backend [60]. We used the Adam optimizer. We trained three networks with the same hyperparameters: batch size = 1, learning rate =  $10^{-4}$ , epochs = 10, and  $\alpha = 1$ .

## 4. Results and Discussion

In this experiment, we separately trained the proposed networks with different  $\beta$  values. We optimized the parameters by the validation set and reported results in our test set. The predicted deformation field could not guarantee diffeomorphism; therefore, the transformation of irreversible regions caused an image to “fold” on itself. In these regions, the determinant of the Jacobian matrix of the deformation field was negative (Figure 4). However, spatial folding is physically impossible; hence, this phenomenon causes registration errors in clinical applications. The frequency of such errors limits the application of neural networks in image registration.

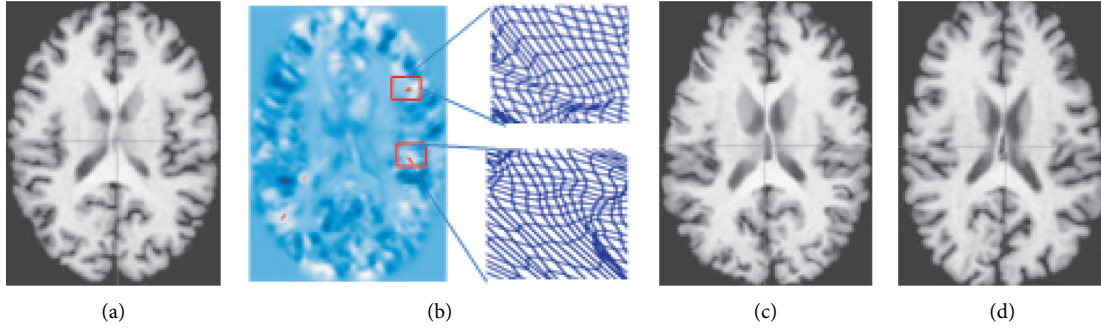


FIGURE 4: The first and last images are the source and target images, respectively, and the third image is the deformed image produced by the method. The second image shows the values of the Jacobian determinant of the predicted deformation with “folding” locations (negative determinant) marked in red. The deformed grids illustrate parts of the deformation. (a) Source, (b) deformation, (c) deformed, and (d) target.

**4.1. Dice Scores.** The mean Dice scores of the different methods across all predicted labels with their corresponding target labels are shown in Table 2. We selected five scales of regularization strength  $\beta$  from 0 to  $10^{-2}$ . Results showed that FSL was not suitable for fine registration because of its few registration parameters in the affine registration with 12 degrees of freedom. ANT (SyN) is a nonrigid registration method, and its registration accuracy was found to be higher than that of affine registration. The Dice scores of FAIM slightly decreased as  $\beta$  increased, and its Dice scores were higher than those of VoxelMorph. The registration accuracy of Probab-Mul was slightly better than that of FAIM. The proposed method achieved the highest registration accuracy under all  $\beta$  values.

The results of the ablation study revealed that the Dice score of Our-GN was higher than FAIM by 2.8% on average (Table 2). Experimental results showed that GN could improve the accuracy of registration. Moreover, the registration accuracy of Our-DO was slightly lower than that of FAIM (Table 2), illustrating that adding a Dropout layer had little impact on registration accuracy.

When  $\beta$  takes 0,  $10^{-5}$ ,  $10^{-4}$ ,  $10^{-3}$ , and  $10^{-2}$ , the Dice scores of our method were higher than those of FAIM by 2.63%, 2.80%, 2.84%, 2.50%, and 3.37%, respectively, higher than those of Probab-Mul by 2.31%, 2.49%, 2.55%, 2.05%, and 2.88%, respectively, and higher than those of VoxelMorph by 4.78%, 5.17%, 5.27%, 5.21%, and 5.90%, respectively (Table 2). This result implied that inserting GN layers into the network architecture could indeed enable the network to learn better parameters and could make the network easier to optimize and converge, thereby improving registration accuracy. During training, we set epochs = 10, and each epoch took about 100 min to perform 2900 iterations.

Figure 5 presents the boxplot of the Dice scores of the five main anatomical regions of the cerebral cortex when  $\beta = 10^{-3}$ . The Dice scores of ANTs in each label were quite different, indicating that ANTs were unstable. The flatness of the boxplots indicated that the stability of the proposed method was comparable to that of other deep learning methods. The proposed method achieved the highest registration accuracy in the five regions of the cerebral cortex.

TABLE 2: Mean Dice scores with different  $\beta$  values.

Mean Dice	$\beta = 0$	$\beta = 10^{-5}$	$\beta = 10^{-4}$	$\beta = 10^{-3}$	$\beta = 10^{-2}$
FSL (Affine)	0.4357	—	—	—	—
ANTs(SyN)	0.5139	—	—	—	—
VoxelMorph	0.5255	0.5203	0.5165	0.5091	0.4908
FAIM	0.5470	0.5440	0.5408	0.5362	0.5161
Probab-Mul	0.5502	0.5471	0.5437	0.5407	0.5210
Our-DO	0.5459	0.5421	0.5380	0.5323	0.5149
Our-GN	0.5729	0.5709	0.5679	0.5591	0.5410
Our method	<b>0.5733</b>	<b>0.5720</b>	<b>0.5692</b>	<b>0.5612</b>	<b>0.5498</b>

Bold values mean the optimal dice score of all methods at the same  $\beta$  value.

Figure 6 shows the mean Dice scores corresponding to the different  $\beta$  values of all methods. The accuracy of the proposed method was relatively consistent with different  $\beta$  values and was higher than that of the other methods.

**4.2. Data-Regularized Penalty Folding.** Figure 7 visualizes the effects of the second regularization term  $R_2(u)$ , which directly penalized “foldings” during training.  $\beta = 0$  means the regularization was not used, and multiple locations were visible in the transformation whose Jacobian determinants were negative. The number of “foldings” voxels greatly reduced when  $\beta = 10^{-5}$ . Only several “folding” voxels were observed when  $\beta = 10^{-4}$ . The number of “folding” voxels was almost eliminated when  $\beta = 10^{-3}$ .

We listed the mean values of the number of voxels of different methods whose Jacobian determinants were negative in Table 3. The proposed method had a lower number of “foldings” in the predicted deformations as  $\beta$  increased.

**4.3. Uncertainty Measure.** In this section, the performance of the registration model in estimating uncertainty was evaluated. Figure 8 shows the results of uncertainty evaluation by the proposed method. In the experiment, the Dropout rate of the Dropout layer was set to 0.5. During the test, the Dropout layer was always on, and 48 Monte Carlo samplings were performed. The threshold  $T$  of the uncertainty map had an impact on the uncertainty measure. We set the threshold between 0 and 1 with an interval of 0.1. As the threshold increased, the proportion of the uncertain part

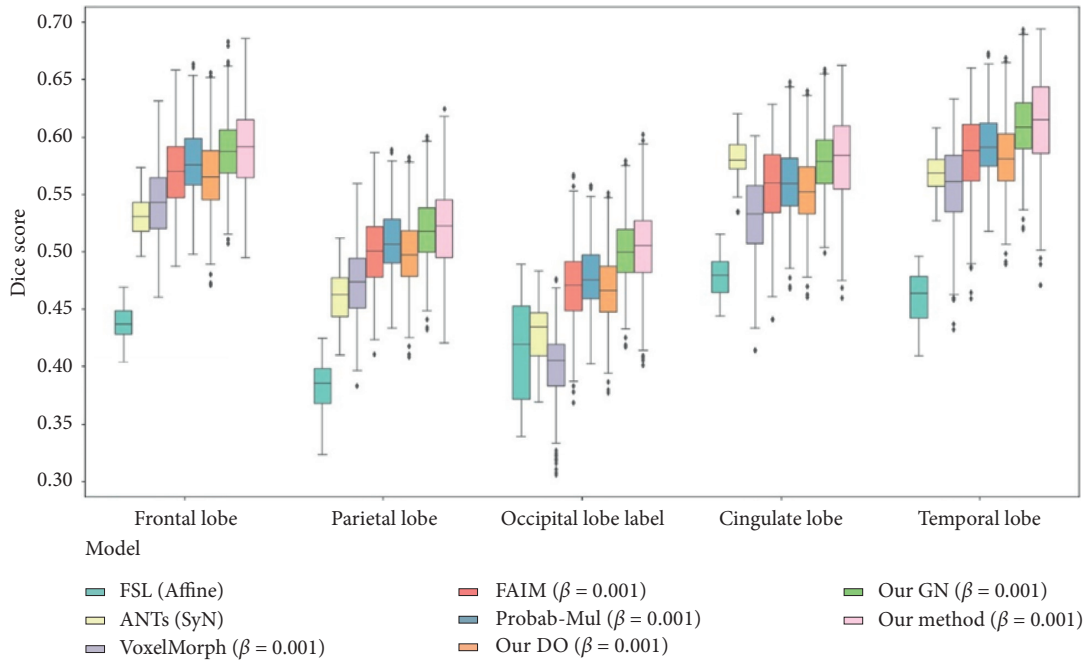


FIGURE 5: Boxplots of Dice scores for five main anatomical structures of the cerebral cortex.

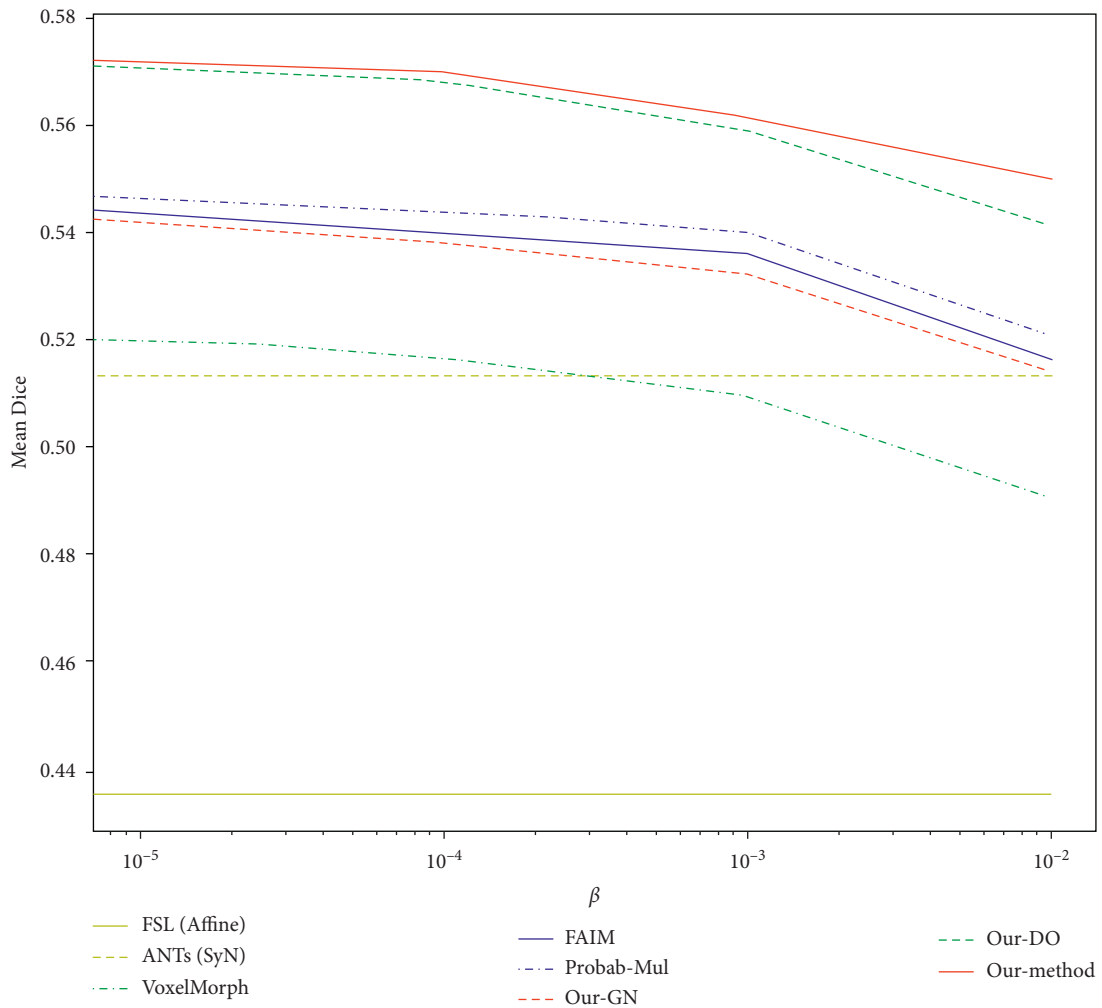


FIGURE 6: Mean Dice score corresponding to the different  $\beta$  of all methods.

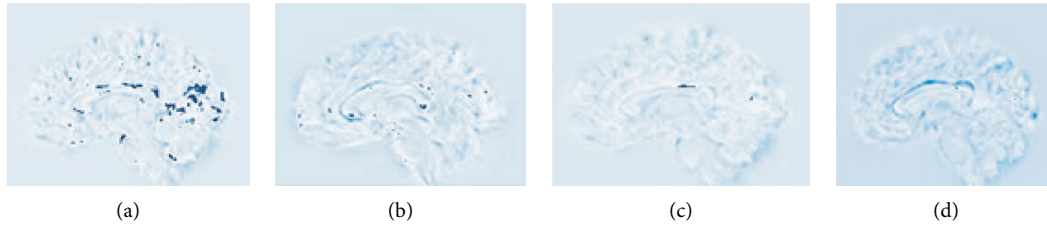


FIGURE 7: Locations where  $\det(\nabla\phi) < 0$  (marked in dark blue) with different  $\beta$  shown on one slice. Predictions were done using the proposed method. (a) 0, (b)  $\beta = 10^{-5}$ , (c)  $\beta = 10^{-4}$ , and (d)  $\beta = 10^{-3}$ .

TABLE 3: Mean number of “folding” locations with different  $\beta$  values.

Mean $N$	$\beta = 0$	$\beta = 10^{-5}$	$\beta = 10^{-4}$	$\beta = 10^{-3}$	$\beta = 10^{-2}$
VoxelMorph	33733	1400	232	60	13
FAIM	39377	1531	234	26	3
Probab-Mul	39905	1700	241	28	6
Our method	39842	1680	240	25	3

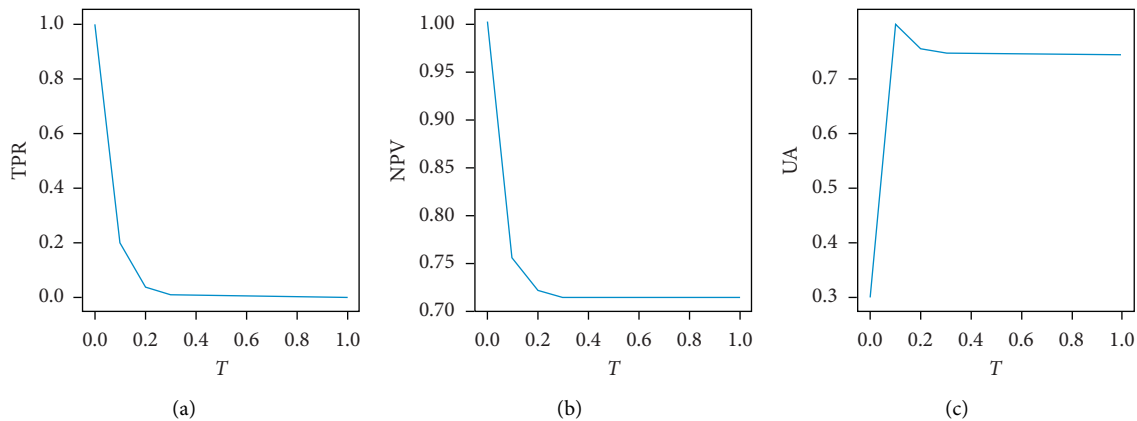


FIGURE 8: Quantitative uncertainty estimation performance for registration task using the evaluation metrics. The abscissa is the threshold ( $T$ ). The ordinate is negative predictive value (NPV), true positive rate (TPR), and uncertainty accuracy (UA), respectively.

of the uncertainty map decreased; hence, the TPR curve gradually decreased. The maximum value of 0.756 in the NPV measurement was obtained at the threshold of 0.1. This value slightly decreased as the threshold further increased but still greater than 0.72. If the model was certain about its prediction, then the accuracy of the prediction was higher. Uncertainty accuracy was also the largest (0.77) when the threshold was 0.1 and slightly declined as the threshold further increased. It remained greater than 0.712. Uncertainty accuracy is the overall accuracy of uncertainty measurements. It shows the ratio of the cases we desired in all possible cases. The uncertainty accuracy of our model was relatively high (Figure 8).

## 5. Conclusion

We developed an unsupervised 3D medical image registration method that uses Bayesian fully convolutional networks for registration. The proposed method introduces probability distributions for network weights and obtains the uncertainty of registration results. We introduced GN into the neural network architecture, which is conducive to

the optimization and convergence of the neural network. The experimental results showed that the proposed method can obtain higher registration Dice scores than other state-of-the-art models and achieve an antifolding performance comparable to that of FAIM and VoxelMorph. The proposed method can also estimate the uncertainty of registration results. Although penalty folding can reduce the irreversible area of registration result, it cannot guarantee that the irreversible area is zero. Thus, the nonrigid registration of diffeomorphism with high accuracy is one of our research directions in the future.

## Data Availability

All datasets used to support the findings of this study were supplied by the publicly available Mindboggle101 database. The URL to access this data is <https://osf.io/yhkde/>.

## Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this work.



## Acknowledgments

The authors thank the National Supercomputing Center in Zhengzhou for the computing resources they provided. This study was funded by the National Natural Science Foundation of China under Grant no. 81772009, Scientific and Technological Research Project of Henan Province under Grant no. 182102310162, Key Research Project of Education Department of Henan Province under Grant no. 21A520042, the Research Fund for Young Scholars of Zhengzhou University under Grant no. F0001297, and Collaborative Innovation Major Project of Zhengzhou under Grant nos. 20XTZX06013 and 20XTZX05015.

## References

- [1] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003.
- [2] M. L. Uss, B. Vozel, S. K. Abramov, and K. Chehdi, "Selection of a similarity measure combination for a wide range of multimodal image registration cases," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 99, pp. 1–16, 2002.
- [3] A. D. Leow, I. Yanovsky, N. Parikshak et al., "Alzheimer's disease neuroimaging initiative: a one-year follow up study using tensor-based morphometry correlating degenerative rates, biomarkers and cognition," *NeuroImage*, vol. 45, no. 3, pp. 645–655, 2019.
- [4] X.-F. Hu, J.-Q. Zhang, X.-M. Jiang et al., "Amplitude of low-frequency oscillations in Parkinson's disease," *Chinese Medical Journal*, vol. 128, no. 5, pp. 593–601, 2015.
- [5] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, "A review of 3D/2D registration methods for image-guided interventions," *Medical Image Analysis*, vol. 16, no. 3, pp. 642–661, 2012.
- [6] Y. Otake, M. Armand, R. S. Armiger et al., "Intraoperative image-based multiview 2D/3D registration for image-guided orthopaedic surgery: incorporation of fiducial-based C-arm tracking and GPU-acceleration," *IEEE Transactions on Medical Imaging*, vol. 31, no. 4, pp. 948–962, 2012.
- [7] M. Ferrant, "Serial registration of intraoperative MR images of the brain," *Medical Image Analysis*, vol. 6, no. 4, pp. 337–359, 2002.
- [8] I. Yanovsky, C. L. Guyader, A. Leow, P. Thompson, and L. Vese, "Unbiased volumetric registration via nonlinear elastic regularization," in *Proceedings of the Miccai Workshop on Mathematical Foundations of Computational Anatomy*, New York, NY, USA, October 2008.
- [9] C. Le Guyader and L. A. Vese, "A combined segmentation and registration framework with a nonlinear elasticity smoother," *Computer Vision and Image Understanding*, vol. 115, no. 12, pp. 1689–1709, 2011.
- [10] C. Davatzikos, "Spatial transformation and registration of brain images using elastically deformable models," *Computer Vision and Image Understanding*, vol. 66, no. 2, pp. 207–222, 1997.
- [11] Y. Wang and L. H. Staib, "Physical model-based non-rigid registration incorporating statistical shape information," *Medical Image Analysis*, vol. 4, no. 1, pp. 7–20, 2000.
- [12] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "A viscous fluid model for multimodal non-rigid image registration using mutual information," *Medical Image Analysis*, vol. 7, no. 4, pp. 565–575, 2003.
- [13] M. C. Chiang, D. Leow, D. Klunder et al., "Fluid registration of diffusion tensor images using information theory," *IEEE Transactions on Medical Imaging*, vol. 7, no. 4, pp. 42–456, 2008.
- [14] H. Lombaert, L. Grady, X. Pennec, N. Ayache, and F. Chriet, "Spectral log-demons: diffeomorphic image registration with very large deformations," *International Journal of Computer Vision*, vol. 107, no. 3, pp. 254–271, 2014.
- [15] N. D. Cahill, J. A. Noble, and D. J. Hawkes, "Demons algorithms for fluid and curvature registration," in *Proceedings of the IEEE International Symposium on Biomedical Imaging: from Nano to Macro*, Boston, MA, USA, July 2009.
- [16] J. Ashburner and K. J. Friston, "Voxel-based morphometry-the methods," *NeuroImage*, vol. 11, no. 6, pp. 805–821, 2000.
- [17] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, and M. O. Leach, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 2002.
- [18] A. V. Dalca, A. Bobu, N. S. Rost, and P. Golland, "Patch-based discrete registration of clinical brain images," in *Proceedings of the International Workshop on Patch-based Techniques in Medical Imaging*, Athens, Greece, October 2016.
- [19] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios, "Dense image registration through MRFs and efficient linear programming," *Medical Image Analysis*, vol. 12, no. 6, pp. 731–741, 2008.
- [20] X. Pennec, P. Cachier, and N. Ayache, "Understanding the 'demon's algorithm': 3D non-rigid registration by gradient descent," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Cambridge, UK, September 1999.
- [21] H. J. Johnson and G. E. Christensen, "Consistent landmark and intensity-based image registration," *IEEE Transactions on Medical Imaging*, vol. 21, no. 5, pp. 450–461, 2002.
- [22] G. Postelnicu, L. Zollei, and B. Fischl, "Combined volumetric and surface registration," *IEEE Transactions on Medical Imaging*, vol. 28, no. 4, pp. 508–522, 2019.
- [23] A. A. Joshi, D. W. Shattuck, P. M. Thompson, and R. M. Leahy, "Surface-constrained volumetric brain registration using harmonic mappings," *IEEE Transactions on Medical Imaging*, vol. 26, no. 12, pp. 1657–1669, 2007.
- [24] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Information measures in medical image registration," *IEEE Transactions on Medical Imaging*, vol. 23, no. 12, pp. 1508–1516, 2004.
- [25] R. Gan, A. Chung, and S. Liao, "Maximum distance-gradient for robust image registration," *Medical Image Analysis*, vol. 12, no. 4, pp. 452–468, 2008.
- [26] M. C. Seiler and F. A. Seiler, "Numerical recipes in C: the art of scientific computing," *Risk Analysis*, vol. 9, no. 3, pp. 415–416, 1989.
- [27] R. H. Byrd, J. Nocedal, and Y. X. Yuan, "Global convergence of a CASS of Quasi-Newton methods on convex problems," *SIAM Journal on Numerical Analysis*, vol. 24, no. 5, pp. 171–1190, 1987.
- [28] B. T. T. Yeo, T. Vercauteren, P. Fillard, J. M. Peyrat, and O. Clatz, "DT-REF in D: diffusion tensor registration with exact finite-strain differential," *IEEE Transactions on Medical Imaging*, vol. 28, no. 12, pp. 1914–1928, 2019.
- [29] B. T. T. Yeo, M. R. Sabuncu, T. Vercauteren, N. Ayache, B. Fischl, and P. Golland, "Spherical demons: fast diffeomorphic landmark-free surface registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 3, pp. 650–668, 2010.



- [30] P. Unser and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2083–2099, 2000.
- [31] S. Klein, J. P. W. Pluim, M. Staring, and M. A. Viergever, "Adaptive stochastic gradient descent optimisation for image registration," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 227–239, 2009.
- [32] S. Ekström, F. Malmberg, H. Ahlström, J. Kullberg, and R. Strand, "Fast graph-cut based optimization for practical dense deformable registration of volume images," *Computerized Medical Imaging and Graphics*, vol. 84, Article ID 101745, 2020.
- [33] B. Avants, C. Epstein, M. Grossman, and J. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.
- [34] H. Y. Zhang, J. W. Zhang, and J. Z. Sun, "Registration method for CT-MR image based on mutual information," *Transactions of Tianjin University*, vol. 13, no. 3, pp. 226–230, 2007.
- [35] V. Villena-Martinez, S. Oprea, and M. Saval-Calvoet, "When deep learning meets data alignment: a review on deep registration networks (DRNs)," 2020, <https://arxiv.org/abs/2003.03167>.
- [36] K. A. J. Eppenhof and J. P. W. Pluim, "Pulmonary CT registration through supervised learning with convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1097–1105, 2019.
- [37] B. Wang, Y. Lei, S. Tian et al., "Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation," *Medical Physics*, vol. 46, no. 4, pp. 1707–1718, 2019.
- [38] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.
- [39] V. Kearney, S. Haaf, A. Sudhyadhom, G. Valdes, and T. D. Solberg, "An unsupervised convolutional neural network-based algorithm for deformable image registration," *Physics in Medicine and Biology*, vol. 63, no. 18, Article ID 185017, 2018.
- [40] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1505–1516, 2016.
- [41] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: a learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.
- [42] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical Image Analysis*, vol. 57, pp. 226–236, 2019.
- [43] D. Kuang and T. Schmäh, "FAIM-a ConvNet method for unsupervised 3D medical image registration," *Machine Learning in Medical Imaging*, Springer, Berlin, Germany, 2019.
- [44] L. H. Liu, X. W. Hu, L. Zhu, and P. A. Heng, "Probabilistic multilayer regularization network for Unsupervised3D brain image registration," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention-MICCAI 2019*, Shenzhen, China, October 2019.
- [45] V. S. Deshpande and J. S. Bhatt, "Bayesian deep learning for deformable medical image registration," in *Proceedings of the International Conference on Pattern Recognition and Machine Intelligence*, Tezpur, India, December 2019.
- [46] S. Khawaled and M. Freiman, "Unsupervised deep-learning based deformable image registration: a Bayesian framework," 2020, <https://arxiv.org/abs/2008.03949>.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 2015.
- [48] Y. Wu and K. He, "Group normalization," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 742–755, 2020.
- [49] R. M. Neal, "Bayesian learning for neural networks," *Lectures Notes in Statistics*, Springer, New York, NY, USA, 1996.
- [50] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, pp. 5574–5584, Curran Associates, Inc., New York, NY, USA, 2017.
- [51] A. Mobiny, H. V. Nguyen, S. Moulik, N. Garg, and C. C. Wu, "DropConnect is effective in modeling uncertainty of Bayesian deep networks," 2019, <https://arxiv.org/abs/1906.04569>.
- [52] D. J. C. Mackay, "Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 469–505, 1995.
- [53] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," 2015, <https://arxiv.org/abs/1505.05424>.
- [54] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," 2015, <https://arxiv.org/abs/1603.04467>.
- [55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from over-fitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [56] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," 2015, <https://arxiv.org/abs/1506.02142>.
- [57] K. Arno and T. Jason, "101 labeled brain images and a consistent human cortical labeling protocol," *Frontiers in Neuroence*, vol. 6, no. 171, p. 171, 2012.
- [58] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, 2011.
- [59] A. Rayne, "Keras," *The School Librarian*, vol. 24, no. 8, pp. 259–261, 2013.
- [60] M. Abadi, A. Agarwal, P. Barham et al., "TensorFlow: large-scale machine learning on heterogeneous distributed systems," 2016, <https://arxiv.org/abs/1603.04467>.

## Research Article

# A Semiautomated Deep Learning Approach for Pancreas Segmentation

Meixiang Huang <sup>1</sup>, Chongfei Huang,<sup>1</sup> Jing Yuan,<sup>2</sup> and Dexing Kong <sup>1</sup>

<sup>1</sup>The School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, China

<sup>2</sup>The School of Mathematics and Statistics, Xidian University, Xi'an 710069, China

Correspondence should be addressed to Dexing Kong; dxkong@zju.edu.cn

Received 25 April 2021; Revised 28 May 2021; Accepted 21 June 2021; Published 3 July 2021

Academic Editor: Jialin Peng

Copyright © 2021 Meixiang Huang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Accurate pancreas segmentation from 3D CT volumes is important for pancreas diseases therapy. It is challenging to accurately delineate the pancreas due to the poor intensity contrast and intrinsic large variations in volume, shape, and location. In this paper, we propose a semiautomated deformable U-Net, i.e., DUNet for the pancreas segmentation. The key innovation of our proposed method is a deformable convolution module, which adaptively adds learned offsets to each sampling position of 2D convolutional kernel to enhance feature representation. Combining deformable convolution module with U-Net enables our DUNet to flexibly capture pancreatic features and improve the geometric modeling capability of U-Net. Moreover, a nonlinear Dice-based loss function is designed to tackle the class-imbalanced problem in the pancreas segmentation. Experimental results show that our proposed method outperforms all comparison methods on the same NIH dataset.

## 1. Introduction

Pancreatic diseases are relatively hidden and difficult to detect and cure, especially for pancreatic cancers, which have high mortality rate worldwide [1]. Accurate pancreas segmentation from 3D CT scans can provide assistance to doctors in the diagnosis of pancreas diseases, such as volumetric measurement and analysis for diabetic patients, as well as surgical guidance for clinicians [2]. However, it is challenging to segment the pancreas due to the large anatomical variability in pancreas position, size, and shape across patients (as shown in Figure 1). Moreover, the ambiguous boundaries around the pancreas with its adjacent structures further increase the difficulty of pancreas delineation.

Traditional methods on abdominal pancreas segmentation mainly have statistical shape models [3, 4] or multi-atlas techniques [5, 6]. Wolz et al. proposed a fully automated method based on a hierarchical atlas registration and weighting scheme for abdominal multiorgan segmentation [6]. This method was evaluated on a database of 150 CT scans and achieved Dice score of 70% for the pancreas. Karasawa

et al. exploited the vasculature around the pancreas to better select atlases for pancreas segmentation [7]. This method was evaluated on 150 abdominal CT scans and obtained an average Dice score of 78.5%. However, the performance of atlas-based methods highly relies on the selection of atlases and the accuracy of the image registration algorithm. Above all, it is difficult to select atlases that are general enough to cover all variabilities in the pancreas across different patients.

Convolutional networks [8, 9] have achieved great success in medical image segmentation, which also boost the performance of pancreas segmentation. U-Net [10], a semantic segmentation architecture, attracted great attentions from researchers by exploiting multilevel feature fusion. The skip connections in U-Net are used to incorporate high-resolution low-level feature maps from the encoding branch into the decoding branch of U-Net to alleviate the important information loss caused by successive downsampling and then refine and recover target details. Namely, using skip connections to fuse multilevel feature tensors can effectively localize and segment target organs [11]. Many works [12–14] have demonstrated that U-Net is a good framework for

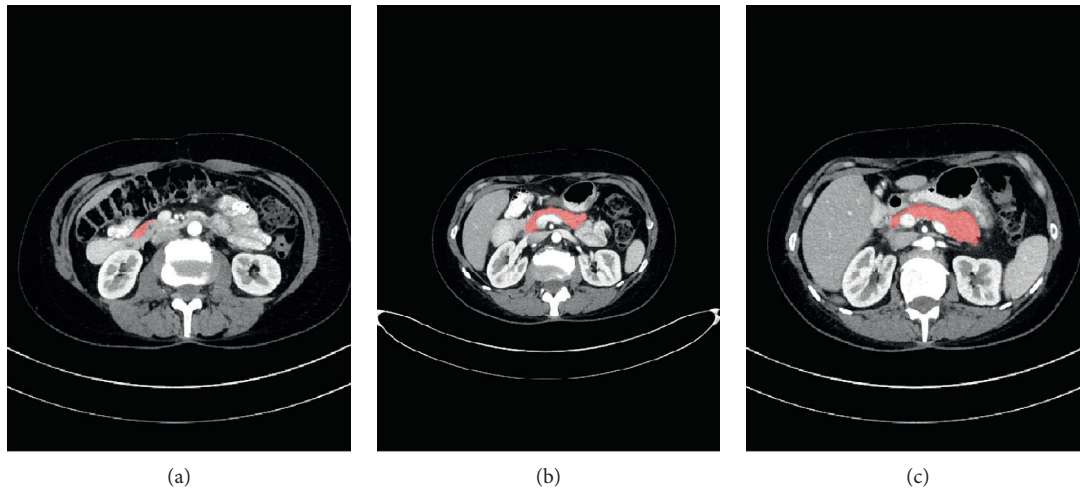


FIGURE 1: Examples of 2D CT slices with pancreas annotations (red regions), showing the highly variable shape and size of pancreas. The largest area of pancreas is less than 0.8% of entire slice while the smallest area is less than 0.1% (best viewed in color).

semantic segmentation tasks, especially for small datasets. Since the pancreas is a small, soft organ in the abdomen, most pancreas segmentation algorithms based on convolutional neural network (CNN) provide iterative algorithms [15] in a coarse-to-fine manner to relieve the interference of complex background. Roth et al. first proposed a probabilistic bottom-up, coarse-to-fine approach for pancreas segmentation [16] where a multilevel deep ConvNet model is utilized to learn robust pancreas features. Two subsequent holistically nested segmentation networks [17, 18] advanced this previous work [16]. Zhou et al. presented a two-stage, fixed-point approach for the pancreas segmentation, which utilized the predicted segmentations from coarse model to localize and obtain smaller pancreas regions, which were further refined by another model [14]. Yu et al. presented the recurrent saliency transformation network to tackle the challenge of small organ segmentation where a saliency transformation module is utilized to connect coarse and fine stage to realize joint optimization [19]. Cai et al. designed a convolutional neural network equipped with convolutional LSTM to impose spatial contextual consistency constraints on successive image slices [20]. Cai et al. [21] further improved the pancreas initial segmentation in [20] by aggregating the multiscale, low-level features and strengthened the pancreatic shape continuity by bidirectional recurrent neural network (BiRNN). Liu et al. [22] used superpixel-based approach to obtain coarse pancreas segmentations, which were then used to train five same-architecture fully convolutional networks (FCNs) with different loss functions to achieve accurate pancreas segmentations. This method is evaluated on 82 public CT volumes and achieved a Dice coefficient of  $84.10 \pm 4.91\%$ . Man et al. [23] proposed a two-stage method composed of deep Q network (DQN) and deformable U-Net for the pancreas segmentation, in which DQN is used to obtain context-adaptive, coarse pancreas segmentations, which were then input to deformable U-Net for refinement. Zhu et al. [24] proposed a 3D coarse-to-fine network to segment the pancreas. This 3D method outperformed the 2D

counterpart due to the full usage of the rich spatial information along the long axial dimension. Some common techniques such as dense connection [25], residual block, and sparse convolution [26, 27] are also widely utilized to segment the pancreas.

Google DeepMind proposed a spatial transformer [28], which is the first work to allow neural networks learn the transformation matrix from data and transform feature maps spatially. Specifically, spatial transformer network (STN) can globally deform feature maps through learned transformations, such as scaling, cropping, rotation as well as nonrigid deformation. Recently, Dai et al. proposed a deformable convolution to get over the limitation of fixed receptive field in standard convolution [29]. In detail, convolutional kernel with explicit offsets learned from the previous feature maps can adaptively change predefined receptive field in order to extract more target features. The specific deformable convolution is shown in Figure 2, in which some standard convolution layers are first utilized to learn and regress the deformation displacements for each sampling point in the image, and then the learned displacements are added to original sampling positions of the 2D convolution to enable network extract relevant and rich features far from original fixed neighborhood [30]. Different from STN [28], deformable convolution adopts a local and dense, instead of global manner to warp feature maps. Moreover, deformable convolution focuses on learning explicit offset for each neuron instead of kernel weights. Since the pancreas has various scales and shapes across patients and traditional convolutional kernel cannot address well on organs with high deformation due to the fixed receptive field, we believe deformable convolution is more suitable for the task of pancreas segmentation [31].

In this paper, we propose a semiautomated deformable U-Net model utilizing the power of U-Net and Deformable-ConvNets. The proposed architecture for pancreas segmentation has two merits. First, deep segmentation networks such as FCN [9], U-Net [10], and DeepLab [32] easily suffer from confusion by the large, irrelevant background

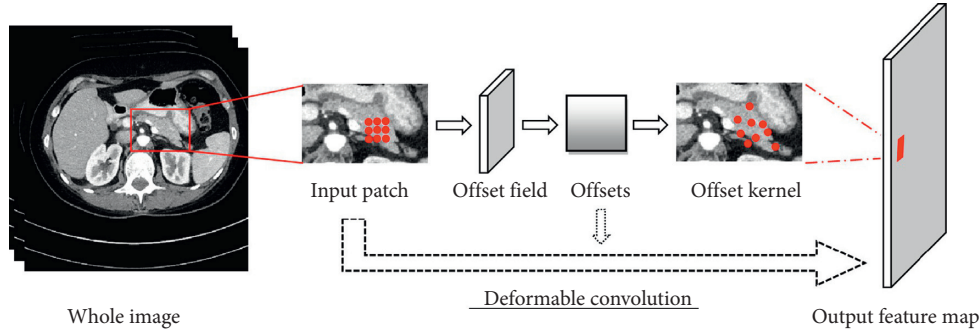


FIGURE 2: Illustration of  $3 \times 3$  deformable convolution. Offset field is generated from the preceding feature maps, and the number of output channels is  $2N$ .

information due to the small size of the pancreas in the entire abdominal CT volume. Motivated by [14], we take a similar strategy, i.e., first manually shrink the size of input image and then refine the extracted pancreas regions by the proposed deformable U-Net. The proposed method has the capability to extract the geometry-aware features of the pancreas with the help of deformable convolution. Second, we propose a novel loss function, focal generalized Dice loss (FGDL) function, to balance the size of foreground and background and enhance the ability of network for small organ segmentation. A conference version of this work was published in ISICDM 2019 [33]. In this extended version, we provide a more comprehensive description of literature review and detailed analysis of the proposed method and experimental investigation. The main modifications include presenting and analyzing the difference between standard convolution block and deformable convolution block (as shown in Figure 3), adding and analyzing the visualization results of the proposed DUNet (as shown in Figures 4 and 5), as well as the comparison results between the proposed DUNet and two baseline methods on the NIH dataset [34] (as shown in Figure 6 and Table 1), adding more evaluation metrics for testing the performance of the proposed DUNet (as shown in (9)–(11)), conducting new experiment to demonstrate the effectiveness of the proposed loss function for pancreas segmentation (as shown in Table 2), discussing advantages and limitations of the proposed DUNet, and adding more references.

## 2. Materials and Methods

In this section, a semiautomated deformable U-Net is proposed to segment the pancreas. Our method is built upon U-Net, which employed skip connections to aggregate multiple feature maps with the same resolution from different levels to recover the grained details lost in decoder branch and thus strengthen the representative capability of network. Since the pancreas only occupies a small fraction of the whole scan and the large and complex background information tends to interfere or confuse semantic segmentation framework, such as U-Net [10], we followed cascade-based methods [5, 12, 14], i.e., first localize target regions and then refine the extracted regions. Specifically, we first estimate the maximum and minimum coordinates of

the pancreas to approximately locate its and then input the extracted pancreas regions to the refinement segmentation model to improve segmentation accuracy. Here, we designed a deformable U-Net (abbreviated as DUNet), as the refinement model. The key component in DUNet is deformable convolution, which can adaptively augment the sampling grid by learning 2D offsets from each image pixel according to the preceding feature maps. Incorporating deformable convolution into the baseline U-Net can improve the geometry-aware capability of U-Net. The overall structure of the proposed method is shown in Figure 7.

**2.1. Network Architecture.** Our approach is an encoder-decoder structure, designed for pancreas segmentation. As shown in Figures 7 and 3, the proposed architecture includes the standard convolution block, deformable convolution block, skip connection, downsampling, and upsampling. Considering that the deformable convolution block requires a little more computing resources and the aim of deformable convolution block is to help the network capture low-level, discriminative details at various shapes and scales, in order to balance the efficiency and accuracy, we experimentally apply the deformable convolution in the second and third layers of U-Net. Specifically, we replaced the standard convolution block of the second and third layers in the encoder, as well as the counterpart layers in the decoder with deformable convolution block. Figure 3(b) shows the component of deformable convolution block. Concretely, each deformable convolution block is composed of convolutional offset layer, followed by convolution layer, BN [35], and ReLU layer, in which convolutional offset layer plays an important role in telling U-Net how to deform and sample feature maps [36]. The advantage of deformable convolution block is to utilize changeable receptive fields to effectively learn pancreas features with various shapes and scales.

Here, we describe the standard convolution and deformable convolution in detail. On the one hand, the standard 2D convolution can be seen as the weighed sum over a regular 2D sampling grid with weight  $W$ . For the  $3 \times 3$  sized kernel with the dilation value of 1 (as shown in Figure 8(a)), the sampling grid  $\mathcal{S}$  in standard convolution defines the receptive field size and can be given by

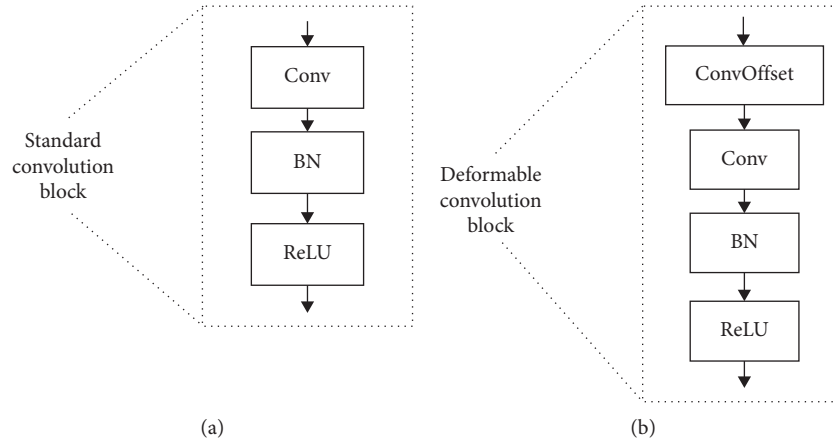


FIGURE 3: The comparison between (a) standard convolution block and (b) deformable convolution block.

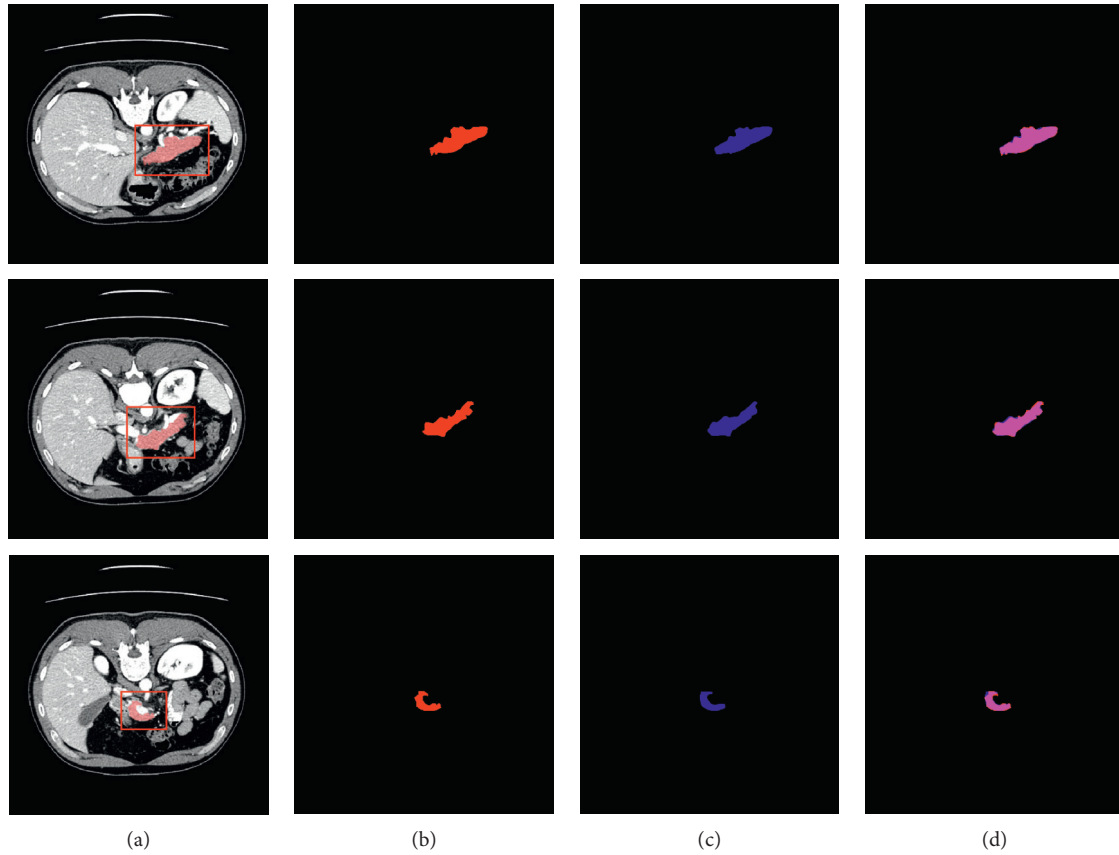


FIGURE 4: Comparisons of 2D pancreas segmentations from the proposed DUNet with the manual segmentations. The first, second, and third columns denote the CT slices with their segmentations and bounding boxes of pancreas (red), the manual segmentations, and the network predictions, respectively. The last column denotes the overlapped maps between the network predictions and manual segmentations, with overlapped regions marked by magenta. (a) Original. (b) Groundtruth. (c) Prediction. (d) Overlapped.

$$\mathcal{E} = \{(-1, -1), (-1, 0), \dots, (0, 1), (1, 1)\}. \quad (1)$$

The value of each location  $p_0$  on the output feature map  $Y$  can be calculated as

$$Y(p_0) = \sum_{p_n \in \mathcal{E}} W(p_n) \cdot X(p_0 + p_n), \quad (2)$$

where  $p_n$  enumerates all locations in 2D sampling grid  $\mathcal{E}$ . On the other hand, rather than using the predefined sampling grid, deformable convolution automatically learns offset  $\Delta p_n$  to augment the regular sampling grid and is calculated as

$$Y(p_0) = \sum_{p_n \in \mathcal{E}} W(p_n) \cdot X(p_0 + p_n + \Delta p_n). \quad (3)$$





FIGURE 5: Comparisons of 3D pancreas segmentations from the proposed DUNet with the manual segmentations. The first, second, and third columns denote the manual segmentations, the network predictions, and the overlapped maps between the network predictions and manual segmentations, respectively. The manual segmentations are shown in red, and the network predictions are shown in light green. (a) Label. (b) Prediction. (c) Overlapped.

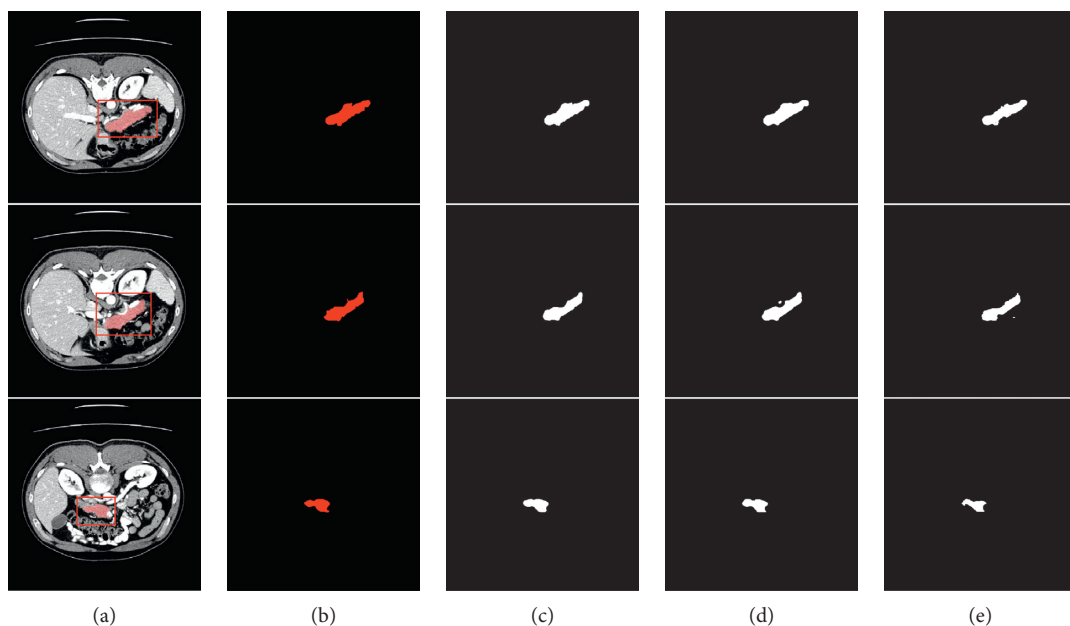


FIGURE 6: Comparison of segmentation results between different models on the NIH dataset. (a) Original images with their segmentations and bounding boxes of pancreas (red). (b) The ground truths. (c–e) The predictions generated by our DUNet, U-Net, and Deformable-ConvNet, respectively.



TABLE 1: Quantitative comparisons between the three different models on the NIH dataset. Bold denotes the best.

Model	F-measure	Recall	Precision	Mean DSC
Modified Deformable-ConvNet	0.8201	0.8084	0.8378	0.8203
U-Net	0.8738	<b>0.9010</b>	0.8499	0.8670
DUNet(Ours)	<b>0.8878</b>	0.8997	<b>0.8898</b>	<b>0.8725</b>

TABLE 2: Comparison of the DUNet with Dice loss (DL) and the proposed loss (DSC%). Bold denotes the best.

Method	Min DSC	Max DSC	Mean DSC
DUNet + DL	68.65	93.18	86.29 $\pm$ 4.33
DUNet + FGDL(Ours)	<b>77.03</b>	<b>93.29</b>	<b>87.25 <math>\pm</math> 3.27</b>

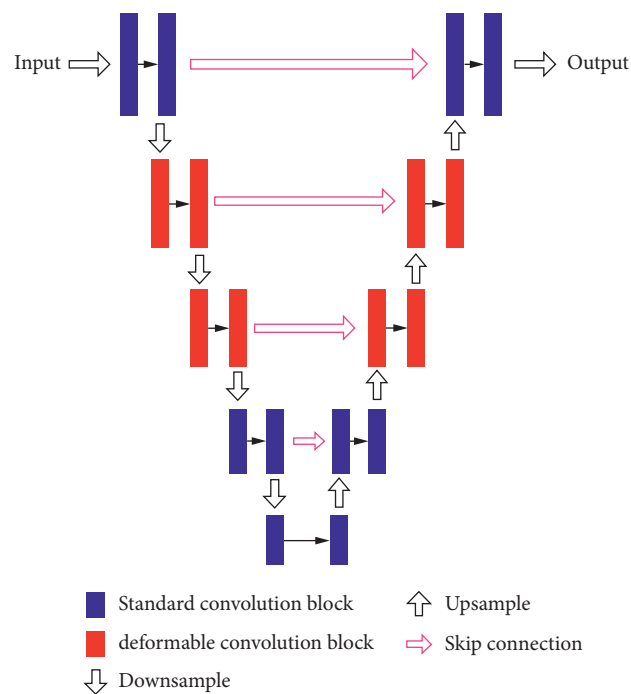
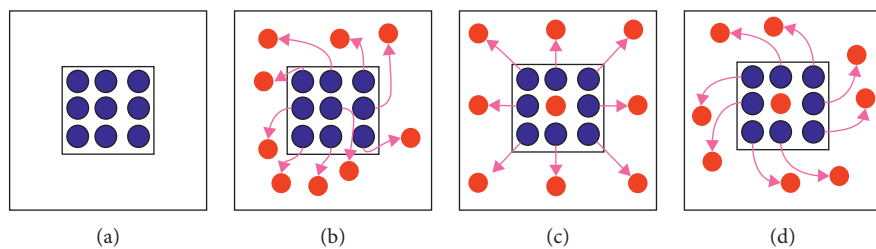


FIGURE 7: An overview of the proposed DUNet. Input data are progressively convolved and downsampled or upsampled by factor of 2 at each scale in both encoding and decoding branches. Schematic of the standard convolution block and deformable convolution block is shown in Figure 3.

FIGURE 8: Comparisons of the sampling points in  $3 \times 3$  standard and deformable convolution. (a) Sampling points (marked as blue) of standard convolution. (b) Deformed sampling points (marked as red) with learned displacements (pink arrows) in deformable convolution. (c-d) Two cases of (b), illustrating that the learned displacements contain translation and rotation transformations.

In particular, the 2D deformable convolution can be mathematically formalized as follows:

$$(W \circ X)(i, j) = \sum_{m=-1}^1 \sum_{n=-1}^1 W(i, j) \times X(i - m + \delta_{i,j,m,n}^{\text{vertical}}, j - n + \delta_{i,j,m,n}^{\text{horizontal}}), \quad \forall i = 1, \dots, H, \forall j = 1, \dots, N, \quad (4)$$

where  $\circ$  denotes the deformable convolution operation,  $W$  is a  $3 \times 3$  kernel with pad 1 and stride 1,  $X$  is the image with height  $H$  and width  $N$ , and  $(i, j)$  denotes the location of pixel in image.  $\delta_{i,j,m,n}^{\text{vertical}}$  and  $\delta_{i,j,m,n}^{\text{horizontal}}$  denote the vertical offset and the horizontal offset, respectively, which are learned by additional convolution on the preceding feature maps. Since the learned offset is usually not an integer, we performed bilinear interpolation on the output of the deformable convolutional layers to enable gradient back-propagation available.

**2.2. Loss Function.** Since the pancreas occupies a small region relative to the large background and Dice loss is relatively insensitive to class-imbalanced problem, most pancreas segmentation works adopt soft, binary Dice loss to optimize pancreas segmentation, and it is defined as follows:

$$L(P, G) = 1 - \frac{\sum_{i=1}^N p_i g_i + \epsilon}{\sum_{i=1}^N p_i + g_i + \epsilon} - \frac{\sum_{i=1}^N (1 - p_i)(1 - g_i) + \epsilon}{\sum_{i=1}^N (2 - p_i - g_i) + \epsilon}, \quad (5)$$

where  $g_i \in \{0, 1\}$  and  $p_i \in [0, 1]$  correspond to the probability value of a voxel in the manual annotation  $G$  and the network prediction  $P$ , respectively.  $N$  and  $\epsilon$  denote the total number of voxels in the image and numerical factor for stable training, respectively. However, Dice loss does not consider the impact of region size on Dice score. To balance the voxel frequency between the foreground and background, Sudre et al. [37] proposed the generalized Dice loss, which is defined as follows:

$$\text{GDL} = 1 - 2 \frac{\sum_{l=1}^2 w_l \sum_i^N p_{li} g_{li}}{\sum_{l=1}^2 w_l \sum_i^N p_{li} + g_{li}}, \quad (6)$$

where coefficient  $w_l = 1/(\sum_{i=1}^N g_{li})$  is a weight for balancing the size of region.

Pancreas boundary plays an important role in delineating the shape of pancreas. However, the pixels around the boundaries of the pancreas are hard samples, which are difficult to delineate due to the ambiguous contrast with the surrounding tissues and organs. Inspired by the focal loss [38, 39], we propose a new loss function, the focal generalized Dice loss (FGDL) function, to alleviate class-imbalanced problem in the pancreas segmentation and allow network to concentrate the learning on those hard samples, such as boundary pixels. The focal generalized Dice loss function can be defined as follows:

$$\text{FGDL} = \sum_{l=1}^2 \left( 1 - 2 \frac{w_l \sum_i^N p_{li} g_{li} + \epsilon}{w_l \sum_i^N p_{li} + g_{li} + \epsilon} \right)^{1/\gamma}, \quad (7)$$

where  $\gamma$  varies in the range  $[1, 3]$ . We experimentally set  $\gamma = 4/3$  during training.

### 3. Experiments

**3.1. Dataset and Evaluation.** We validated the performance of our algorithm on 82 abdominal contrast-enhanced CT images which come from the NIH pancreas segmentation dataset [34]. The original size of each CT scan is  $512 \times 512$  with the number of slices from 181 to 460, as well as the slice thickness from 0.5 mm to 1.0 mm. The image intensity of each scan is truncated to  $[-100, 240]$  HU to filter out the irrelevant details and further normalized to  $[0, 1]$ . In this study, we cropped each slice to  $[192, 256]$ . For fair comparisons, we trained and evaluated the proposed model with 4-fold cross validation.

Four metrics including the Dice Similarity Coefficient (DSC), Precision, Recall, and F-measure (abbreviated as  $F_1$ ) [40] are used to quantitatively evaluate the performance of different methods.

- (1) Dice Similarity Coefficient (DSC) measures the volumetric overlap ratio between the ground truths and network predictions. It is defined as follows [41]:

$$\text{DSC} = \frac{2 \|V_{gt} \cap V_{seg}\|}{\|V_{gt}\| + \|V_{seg}\|}, \quad (8)$$

- (2) Precision measures the proportion of truly positive voxels in the predictions. It is defined as follows:

$$\text{Precision} = \frac{\|V_{gt} \cap V_{seg}\|}{\|V_{seg}\|}. \quad (9)$$

- (3) Recall measures the proportion of positives that are correctly identified. It is defined as follows:

$$\text{Recall} = \frac{\|V_{gt} \cap V_{seg}\|}{\|V_{gt}\|}. \quad (10)$$

- (4) F-measure shows the similarity and diversity of testing data. It is defined as follows:

$$F\text{-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (11)$$

where  $V_{gt}$  and  $V_{seg}$  represent the voxel sets of manual annotations and network predictions, respectively. For DSC, the experimental results are all reported as the mean with standard deviation over all 82 samples. For Precision, Recall, and F-measure metrics, we just reported the mean score over all 82 samples.

TABLE 3: Comparison with other segmentation methods on the NIH dataset (DSC%). Bold denotes the best.

Method	Min DSC	Max DSC	Mean DSC
Roth et al., MICCAI'2015 [16]	23.99	86.29	71.42 $\pm$ 10.11
Roth et al., MICCAI'2016 [17]	34.11	88.65	78.01 $\pm$ 8.20
Zhou et al., MICCAI'2017 [14]	62.43	90.85	82.37 $\pm$ 5.68
Cai et al., 2019 [21]	59.00	91.00	83.70 $\pm$ 5.10
Liu et al., IEEE access 2019 [22]	N/A	N/A	84.10 $\pm$ 4.91
Zhu et al., 3DV'2018 [24]	69.62	91.45	84.59 $\pm$ 4.86
Man et al., IEEE T MED IMAGING 2019 [23]	74.32	91.34	86.93 $\pm$ 4.92
DUNet(Ours)	<b>77.03</b>	<b>93.29</b>	<b>87.25 <math>\pm</math> 3.27</b>

**3.2. Implementation Details.** The proposed method was implemented on the Keras and TensorFlow platforms and trained using Adam optimizers for 10 epochs on a NVIDIA Tesla P40 with 24 GB GPU. The learning rate and batch size were set to 0.0001 and 6 for training, respectively. In total, the trainable parameters in the proposed DUNet are 6.44 M, and the average inference time of our DUNet per volume is 0.143 seconds.

**3.3. Qualitative and Quantitative Segmentation Results.** To assess the effectiveness of deformable convolution in the pancreas segmentation, we compared the three models: Deformable-ConvNet, U-Net, and DUNet. To make the output size of Deformable-ConvNet to be the same as input, we make modification on Deformable-ConvNet [29] by substituting the original fully connected layers with upsampling layers. Figure 6 qualitatively shows the improvements brought by deformable convolution. It can be observed that our DUNet focuses more on the details of the pancreas, which demonstrates that deformable convolution can extract more pancreas information and enhance the geometric recognition capability of U-Net.

The quantitative comparisons of different models in terms of the Precision, Recall,  $F_1$ , and mean DSC are reported in Table 1. It can be observed that our DUNet outperforms the modified Deformable-ConvNet and U-Net with improvements of average DSC up to 5.22% and 0.55%. Furthermore, it is worth noting that our proposed DUNet reported the highest average F-measure with 88.78%, which demonstrates that the proposed DUNet is a high-quality segmentation model and more robust than other two approaches. Figures 4 and 5 visualize the 2D and 3D overlap of segmentations from the proposed DUNet with respect to the manual segmentations, respectively. Visual inspection of the overlapping maps shows that the proposed DUNet can fit the manual segmentations well, which further demonstrates the effectiveness of our method.

**3.4. Impact of Loss Function.** To assess the effectiveness of the proposed loss function, we test standard Dice loss and the proposed loss with DUNet, i.e., Dice loss and the proposed focal generalized Dice loss (FGDL); the segmentation performance of the DUNet with different loss function is reported in Table 2. It can be noted that DUNet with the proposed FGDL improves mean DSC by 0.96% and min DSC by 8.38% compared with Dice loss.

**3.5. Comparison with Other Methods.** We compared the segmentation performance of the proposed DUNet with seven approaches [14, 16, 17, 21–24] on the NIH dataset [34]. Note that the experimental results of other seven methods were obtained directly from their corresponding literatures. As shown in Table 3, our method achieves the min DSC of 77.03%, max DSC of 93.29%, and mean DSC of 87.25  $\pm$  3.27%, which outperforms all comparison methods. Moreover, the proposed DUNet performed the best in terms of both standard deviation and the worst case, which further demonstrates the reliability of our method in clinical applications.

## 4. Discussion

The pancreas is a very important organ in the body, which plays a crucial role in the decomposition and absorption of blood sugar and many nutrients. To handle the challenges of large shape variations and fuzzy boundaries in the pancreas segmentation, we propose a semiautomated DUNet to adaptively learn the intrinsic shape transformations of the pancreas. In fact, DUNet is an extension of U-Net by substituting the standard convolution block of the second and third layers in the encoder and counterpart layers in the decoder of U-Net with deformable convolution. The main advantage of the proposed DUNet is that DUNet utilizes the changeable receptive fields to automatically learn the inherent shape variations of the pancreas, then extract robust features, and thus improve the accuracy of pancreas segmentation.

There are several limitations in this work. First, during data processing, we first need radiologists to approximately annotate the minimum and maximum coordinates of the pancreas in each slice in order to localize it and thus reduce the interference brought by complex background. This work may be laborious. Second, the trainable parameters are relatively excessive. In future work, we will further improve pancreas segmentation performance from two aspects. First, we will explore and adopt attention mechanism to eliminate localization module and construct a lightweight network. Second, we will consider how to fuse prior knowledge (e.g., shape constraint) to the network.

## 5. Conclusions

In this paper, we proposed a semiautomated DUNet to segment the pancreas, especially for the challenging cases with large shape variation. Specifically, the deformable

convolution and U-Net structure are integrated to adaptively capture meaningful and discriminative features. Then, a nonlinear Dice-based loss function is introduced to supervise the DUNet training and enhance the representative capability of DUNet. Experimental results on the NIH dataset show that the proposed DUNet outperforms all the comparison methods.

## Data Availability

Pancreas CT images used in this paper were from a public available pancreas CT dataset, which can be obtained from <http://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU>.

## Disclosure

An earlier version of our study has been presented as a conference paper in the following link: <https://doi.org/10.1145/3364836.3364894>.

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant nos. 12090020 and 12090025) and Zhejiang Provincial Natural Science Foundation of China (Grant no. LSD19H180005).

## References

- [1] P. Ghaneh, E. Costello, and J. P. Neoptolemos, "Biology and management of pancreatic cancer," *Postgraduate Medical Journal*, vol. 84, no. 995, pp. 478–497, 2008.
- [2] S. V. DeSouza, R. G. Singh, H. D. Yoon, R. Murphy, L. D. Plank, and M. S. Petrov, "Pancreas volume in health and disease: a systematic review and meta-analysis," *Expert Review of Gastroenterology & Hepatology*, vol. 12, no. 8, pp. 757–766, 2018.
- [3] J. J. Cerrolaza, R. M. Summers, and M. G. Linguraru, "Soft multi-organ shape models via generalized pca: a general framework," in *Medical Image Computing And Computer-Assisted Intervention* Springer, Berlin, Germany, 2016.
- [4] A. Saito, S. Nawano, and A. Shimizu, "Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs," *Medical Image Analysis*, vol. 28, pp. 46–65, 2016.
- [5] M. Oda, N. Shimizu, H. R. Roth et al., "3D FCN feature driven regression forest-based pancreas localization and segmentation," in *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 222–230, Quebec, Canada, September 2017.
- [6] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1723–1730, 2013.
- [7] K. I. Karasawa, M. Oda, T. Kitasaka et al., "Multi-atlas pancreas segmentation: atlas selection based on vessel structure," *Medical Image Analysis*, vol. 39, pp. 18–28, 2017.
- [8] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 2, pp. 1097–1105, 2012.
- [9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 7–12, pp. 3431–3440, 2015.
- [10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science*, vol. 9351, pp. 234–241, 2015.
- [11] C. Lyu, G. Hu, and D. Wang, "HRED-net: high-resolution encoder-decoder network for fine-grained image segmentation," *IEEE access*, vol. 8, pp. 38210–38220, 2020.
- [12] A. Farag, L. Lu, H. R. Roth, J. Liu, E. Turkbey, and R. M. Summers, "A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 386–399, 2017.
- [13] X. Li, H. Chen, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [14] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal ct scans," *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017*, vol. 10, pp. 693–701, 2017.
- [15] P. J. Hu, X. Li, Y. Tian et al., "Automatic pancreas segmentation in CT images with distance-based saliency-aware DenseASPP network," *IEEE journal of biomedical and health informatics*, vol. 25, no. 5, pp. 1601–1611, 2020.
- [16] H. R. Roth, L. Lu, A. Farag et al., "DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation," in *Proceedings of the Medical Image Computing And Computer Assisted Intervention*, pp. 556–564, Munich, Germany, June 2015.
- [17] H. R. Roth, L. Lu, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested networks for automated pancreas segmentation," *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*, vol. 9901, pp. 451–459, 2016.
- [18] H. R. Roth, L. Lu, N. Lay et al., "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical Image Analysis*, vol. 45, pp. 94–107, 2018.
- [19] Q. Yu, L. Xie, Y. Wang et al., "Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8280–8289, Salt Lake, UT, USA, June 2018.
- [20] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function," in *Medical Image Computing And Computer-Assisted Intervention* Springer, Berlin, Germany, 2017.
- [21] J. Cai, L. Lu, F. Xing, and L. Yang, "Pancreas segmentation in CT and MRI via task-specific network design and recurrent neural contextual learning," in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics* Springer, Berlin, Germany, 2019.
- [22] S. Liu, X. Yuan, R. Hu, S. Liang, and S. Feng, "Automatic pancreas segmentation via coarse location and ensemble learning," *IEEE Access*, vol. 8, pp. 2906–2914, 2019.

- [23] Y. Man, Y. Huang, J. Feng, X. Li, and F. Wu, "Deep Q learning driven CT pancreas segmentation with geometry-aware U-net," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1971–1980, 2019.
- [24] Z. Zhu, Y. Xia, W. Shen, E. Fishman, and A. Yuille, "A 3D coarse-to-fine framework for volumetric medical image segmentation," in *Proceedings of the International Conference on 3D Vision*, pp. 682–690, Verona, Italy, September 2018.
- [25] E. Gibson, F. Giganti, Y. Hu et al., "Towards image-guided pancreas and biliary endoscopy: automatic multi-organ segmentation on abdominal CT with dense dilated networks," *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017*, vol. 10, pp. 728–736, 2017.
- [26] M. P. Heinrich, M. Blendowski, and O. Oktay, "TernaryNet: faster deep model inference without GPUs for medical 3D segmentation using sparse and binary convolutions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 9, pp. 1311–1320, 2018.
- [27] M. P. Heinrich and O. Oktay, "BRIEFnet: deep pancreas segmentation using binary sparse convolutions," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, vol. 435, pp. 329–337, 2017.
- [28] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 2015, pp. 2017–2025, 2015.
- [29] M. F. Dai, H. Qi, Y. Xiong et al., "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773, Venice, Italy, October 2017.
- [30] Y. Wang, J. Yang, L. Wang et al., "Light field image super-resolution using deformable convolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 1057–1071, 2021.
- [31] S. A. Siddiqui, M. I. Malik, S. Agne, A. Dengel, and S. Ahmed, "DeCNT: deep deformable CNN for table detection," *IEEE access*, vol. 6, pp. 74151–74161, 2018.
- [32] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *International Conference on Learning Representations*, vol. 40, 2015.
- [33] M. X. Hunag, C. F. Huang, J. Yuan, and D. X. Kong, "Fixed-point deformable U-net for pancreas CT segmentation," in *Proceedings of the 3rd International Symposium on Image Computing and Digital Medicine*, pp. 283–287, Xian, China, August 2019.
- [34] H. R. Roth, A. Farag, E. B. Turkbey et al., "Data from pancreas-CT," *The Cancer Imaging Archive*, vol. 32, 2016.
- [35] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, July 2015.
- [36] W. Liu, Y. Song, D. Chen et al., "Deformable object tracking with gated fusion," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3766–3777, 2019.
- [37] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 55, pp. 240–248, 2017.
- [38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.
- [39] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention U-net for lesion segmentation," in *Proceedings of the IEEE 16th International Symposium on Biomedical Imaging*, pp. 683–687, Venice, Italy, April 2019.
- [40] N. Lazarevic-Mcmanus, J. R. Renno, D. Makris, and G. A. Jones, "An object-based comparative methodology for motion detection based on the F-Measure," *Computer Vision and Image Understanding*, vol. 111, no. 1, pp. 74–85, 2008.
- [41] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.



## Research Article

# Tic Detection in Tourette Syndrome Patients Based on Unsupervised Visual Feature Learning

Junya Wu <sup>1</sup>, Tianshu Zhou,<sup>2</sup> Yufan Guo,<sup>3</sup> Yu Tian,<sup>1</sup> Yuting Lou,<sup>3</sup> Hua Ru,<sup>2</sup> Jianhua Feng <sup>3</sup>, and Jingsong Li <sup>1,2</sup>

<sup>1</sup>Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, China

<sup>2</sup>Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou 311100, China

<sup>3</sup>Department of Pediatrics, The Second Affiliated Hospital of Zhejiang University School of Medicine, No. 88 Jiefang Road, Hangzhou 310009, China

Correspondence should be addressed to Jianhua Feng; [hzhz87083886@zju.edu.cn](mailto:hzhz87083886@zju.edu.cn) and Jingsong Li; [ljs@zju.edu.cn](mailto:ljs@zju.edu.cn)

Junya Wu and Tianshu Zhou contributed equally to this work.

Received 5 March 2021; Revised 4 May 2021; Accepted 24 May 2021; Published 7 June 2021

Academic Editor: Jialin Peng

Copyright © 2021 Junya Wu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A clinical diagnosis of tic disorder involves several complex processes, among which observation and evaluation of patient behavior usually require considerable time and effective cooperation between the doctor and the patient. The existing assessment scale has been simplified into qualitative and quantitative assessments of movements and sound twitches over a certain period, but it must still be completed manually. Therefore, we attempt to find an automatic method for detecting tic movement to assist in diagnosis and evaluation. Based on real clinical data, we propose a deep learning architecture that combines both unsupervised and supervised learning methods and learns features from videos for tic motion detection. The model is trained using leave-one-subject-out cross-validation for both binary and multiclass classification tasks. For these tasks, the model reaches average recognition precisions of 86.33% and 86.26% and recalls of 77.07% and 78.78%, respectively. The visualization of features learned from the unsupervised stage indicates the distinguishability of the two types of tics and the nontic. Further evaluation results suggest its potential clinical application for auxiliary diagnoses and evaluations of treatment effects.

## 1. Introduction

Tourette syndrome (TS) is a childhood-onset neurodevelopmental disorder characterized by the presence of fluctuating motor and vocal tics [1]. The core diagnostic features are both multiple motor and one or more phonic tics lasting more than one year. Typically, the same tic occurs at short-term periodicity with short intervals [2]. The simple tic forms are eye blinking, mouth twitching, head jerking, etc. Multiple studies published since 2000 have consistently demonstrated that the prevalence of TS is much higher than previously thought [3]. As the understanding of this disease deepens, the number of children diagnosed with tic disorder has gradually increased, but most cases do not receive timely clinical attention in the early

stages of the disease. Furthermore, approximately 20% of persons with TS are unaware that they have tics [4]. The clinical diagnosis of TS involves complex processes that require considerable time and effective cooperation between the doctor and the patient, especially observation and evaluation of the patient's tic behaviors. A number of instruments for tics and associated phenomena have been developed to assess tic severity [5] and differ in construct, comprehensiveness, and ease of administration.

Recently, artificial intelligence and machine learning have been widely applied in the medical field. In particular, the development of video-based human motion behavior analysis technology has advanced various types of medical diagnoses, such as Parkinson's disease [6], seizure disorders

[7], spinal muscular atrophy [8], and discomfort detection in premature infants [9]. In part, noncontact video-based analysis has attracted great attention due to the increasing availability of camera monitoring systems. To identify tic disorders, patients' tic movements can be detected and analyzed from video tapes that show the patient's face, head, or body and rated according to the Yale Global Tic Severity Scale (YGTSS) [10] or the Modified Rush Video-based Rating Scale (MRVRS) [11]. These ratings can then be used to assist a clinical doctor in evaluating the patient's symptoms and severity.

Tic movements can be distributed throughout the body. Rickard's research [12] showed that patients' twitches usually start in the facial area and that eye twitches are the most frequent. Chappell et al. [13] showed that, in addition to a severity scale, the severity of Tourette syndrome can also be determined by recording the patient's tics with video for more than ten minutes. Moreover, monitoring and recording patients in their natural states instead of facing a clinician effectively avoids interference in diagnosis and evaluation caused by the patient actively controlling their tics. Therefore, we aim to develop a method to automatically detect tics to help clinicians or parents spot and assess tic signs.

In recent decades, many studies have focused on the pathology, genetics, and clinical treatment of TS [14–16], but only a few studies have been published regarding the automatic detection of TS-related motor disturbances. Jonathan et al. [17] studied two patients with TS using deep brain stimulation (DBS) during tics and found that low-frequency (1–10 Hz) centromedian (CM) thalamic activity and beta frequency motor cortex (M1) activity were tic features and that long complex tics are concurrent with a highly detectable thalamocortical signature. Bernabei et al. [4] used a wearable device attached to the patient's trunk with an embedded triaxial accelerometer to monitor tic events. This approach achieved a sensitivity of 80.8%, a specificity of 75.8%, and an accuracy of 80.5%. However, the implementation process of this method is quite complicated, which poses a major challenge and requires extensive cooperation between doctors and patients. Recently, Barua et al. [18] proposed a deep learning approach for detecting tic disorders using wireless channel information and achieved an accuracy above 97%. The data used in the task were simulated using healthy human subjects. However, in a real clinical situation, acquiring such data would be a considerably more complicated task. Regarding methodological aspects, action detection methods have made numerous advancements in video comprehension, such as the two-stream network [19–21], 3D ConvNet [22–24], and temporal enhancement-and-interaction network (TEINET) [25], whereas these deep learning networks require large amounts of labeled data, which carries the high costs and slow procedures associated with manual labeling. Data labeling is often costly and time consuming; an example is the popular ImageNet dataset [26]. However, in real-world situations, large amounts of readily accessible unlabeled data exist; therefore, unsupervised learning has attracted increasing attention from researchers.

From these perspectives, we instead adapt a two-stage architecture by first training an unsupervised feature extraction model to make full use of the more easily acquired

unlabeled data and then applying a comparatively simple network attached to the former trained model for the classification tasks. Visualizing the feature representation of the labeled data shows the correspondence with the tic parts, indicating that the unsupervised model learned the valid feature representation. This approach results in the following contributions: (1) we employ a deep learning scheme by a convolutional-neural-network- (CNN-) based model to learn feature representation from abundant unlabeled video data, (2) we apply a long short-term memory neural network (LSTM) to classify the feature sequences of video clips, and (3) an automated video-based system for detecting tic movements in TS patients is devised.

## 2. Materials and Methods

To solve the problem of insufficient labeled data but enough monitoring video data, we propose a two-stage framework that combines unsupervised and supervised learning, as shown in Figure 1. In the first stage, we adopt a contrastive learning network that learns from unlabeled video data by extracting features by maximizing mutual information. The core idea behind this is to maximize the mutual information between the two nonoverlapping patch inputs. In the second stage, we design an end-to-end architecture based on an LSTM network connected to the feature extraction module in the first stage that learns to classify tic movements from video data labeled by doctors. We use a combination of supervised and unsupervised learning to design and build an end-to-end tic detection model.

*2.1. Subjects.* Sixty-eight patients (4–13 years old) diagnosed with TS by two experienced specialists were employed in this study. All participants were inpatients under normal treatment recruited from the Second Affiliated Hospital of Zhejiang University School of Medicine between May and September 2019. This study was approved by the ethics committee of the Second Affiliated Hospital of Zhejiang University School of Medicine (YAN2019-148). All participants provided written informed consent with the agreement to participate in the study.

*2.2. Data Acquisition and Preprocessing.* The TS dataset was sourced from the Department of Pediatrics at the Second Affiliated Hospital of Zhejiang University School of Medicine and was collected using EEG video acquisition equipment installed in the pediatric ward. The video data were recorded in two situations: (i) the patient was asked to sit on a chair in front of the camera and (ii) before or after EEG recording, the patient was asked to agree to video recording. The two situations arise because the data were collected in different periods: (i) represents data collected during the preproject preparation phase, whereas (ii) is a part of the routine during subsequent EEG video recording. In both situations, every patient was informed in advance of the recording period and asked to face the camera as much as possible during recording, but no mandatory measures were imposed; the patients could move freely, which may result in

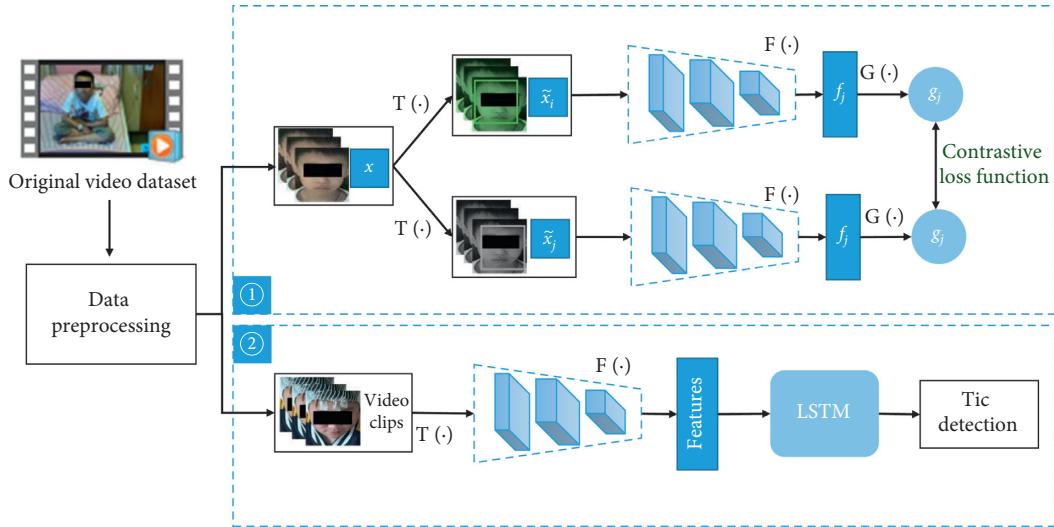


FIGURE 1: The architecture of the proposed method. (1) Stage 1: extracting representative visual features. (2) Stage 2: training an LSTM using visual features.

useless frames. The patients' parents provided informed permission for the collection procedure and research use of the video recordings.

Due to different camera devices, the original video frame rate includes video acquired at both 30 and 25 fps. The duration of all videos ranges from 5 to 15 minutes, and all the videos have a resolution of  $1920 \times 1080$ . The length of the videos is listed in Table 1, and the distribution of the durations is presented in Figure S1. In this TS dataset, 13 cases were annotated by two specialists, who labeled the starting and ending timestamps of a specific tic event, such as an eye tic or a mouth tic. They performed manual annotation frame by frame through the video annotation software VoTT (<https://github.com/microsoft/VoTT>), which then generated annotated JavaScript Object Notation (JSON) files for postprocessing to extract annotations. The annotation work was independently performed by two clinicians and verified by the third expert, and they performed an extra check if there were disagreements until they reached a consensus. Finally, we cut and categorized the videos based on the timestamped annotations to form the TS research dataset, which can be supplemented at any time. We also segmented the video segments between two labeled tic events in the original video, which can be used as normal recordings and act as negative samples.

There are more than five types of tics involving different muscle groups in the labeled data. The most common tics are eye tic and mouth tic. Not only in this research dataset but also most in clinical practice, these two tic types are widespread from a specialist's perspective [12]. Therefore, we defined two classification tasks. (1) We chose these two tic types and normal recordings to define a multiclass classification task. (2) We also configured a binary classification task for the tic and normal datasets—that is, all the tic video clips form the positive samples, while the normal video clips form the negative samples. Figure 2 shows the category proportion of every patient in the labeled dataset, defined in these two tasks.

The proposed method is composed of two stages, and there are slight differences in data preprocessing for the two

TABLE 1: Original TS video dataset.

Category	Labeled dataset	Unlabeled dataset
Videos	13	55
Minutes	136	709

stages. The common operation is to obtain the region of interest (ROI), which is defined as the area centered on the patient's face. This ensures that the models will focus on features related to patients' tic behaviors rather than on other family members or physicians visible in the videos. Identifying the ROI also reduces interference from different camera angles and from patient movements since they are free to move out of the camera view. This procedure uses a neural-network-driven face detection method. We use the multitask cascaded CNN (MTCNN) [27] architecture to detect the patient's face and obtain the ROI and use the pretrained weights from Face2 [28]. To avoid the regional deviations caused by free patient motion and obtain more features in the face area, we extract the ROI area by expanding the width of the face bounding box by 20%. Figure 3 illustrates the ROI output margin. We also conduct data augmentation during preprocessing, an approach that has been widely used in both supervised and unsupervised learning [29, 30]. The effectiveness of simple data augmentation methods for contrastive learning tasks was verified by [31]. Similarly, after obtaining the ROI area, combined data augmentation methods are adopted including random cropping, random noise, and random color distortion.

The difference between the two main stages during data preprocessing is that the first stage uses a relatively large number of frames from an unlabeled video dataset. As well known, the information between continuous video frames is usually highly redundant, which can cause overfitting during training. Thus, we perform a 3-fold downsampling procedure, which extracts the first frame for every three

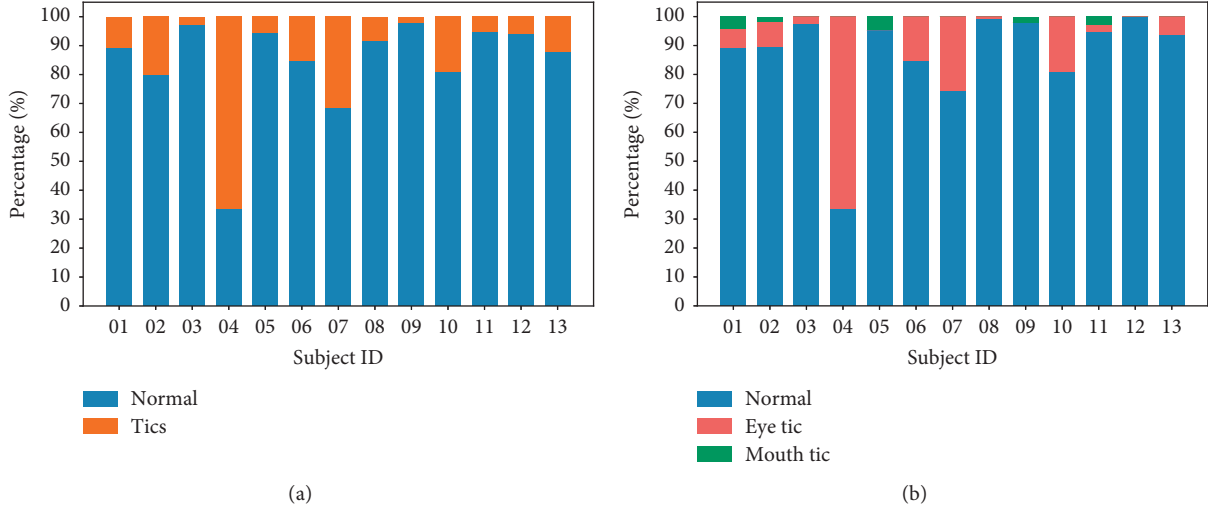


FIGURE 2: Category proportion of normalized video clips of patients in (a) the multiclass classification task and (b) the binary classification task. The ordinate indicates the patient number in the labeled TS dataset.

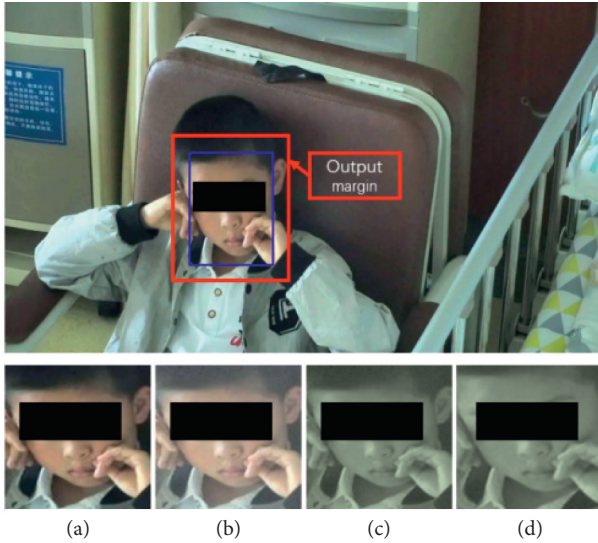


FIGURE 3: Region of interest in data preprocessing. (a) is the output of the MTCNN, and (b), (c), and (d) are the random data augmentation methods applied.

consecutive frames on the original videos. In the second supervised training stage, we first cut the original continuous video data into video segments and divide them into motion tic categories based on annotations. Then, we cut each video segment into 1 s video clips using a no overlapping sliding window. These clips form the input objects for Stage 2. We then performed the same ROI extraction procedure as for Stage 1 but without video frame downsampling because the frame data are randomly extracted from every second of input, which has an effect similar to downsampling.

**2.3. Stage 1: Extracting Representative Visual Features.** The dataset used in this work consists of a small amount of labeled data ( $n=13$ ) and a relatively large amount of

unlabeled data ( $n=55$ ). Apparently, the labeled data we have are insufficient to train a deep learning model. To explore the value of the unlabeled data, we adopt a contrastive learning framework similar to SimCLR [31] in Stage 1 to extract representative visual features among the TS patient groups. Specifically, a randomly selected minibatch  $S$  of  $N$  examples is transformed into a minibatch pair  $S'$  consisting of  $2N$  examples after applying a combination of a set of data augmentation methods. Then,  $S'$  is input to the defined contrastive prediction task. For every minibatch pair  $S'$ , each pair  $(i, j)$  of augmented examples  $S'(i, j)$  is treated as a positive example ( $n=2$ ), while the others ( $n=2(N-1)$ ) are treated as negative examples. Then the similarity  $\text{sim}(i, j)$  of the pair  $S'(i, j)$  is defined as follows:

$$\text{sim}(i, j) = \frac{S'_i{}^T S'_j}{\left(\|S'_i\| \|S'_j\|\right)}, \quad (1)$$

and the loss function of the pair  $\text{loss}(i, j)$  is defined as

$$\text{loss}(i, j) = -\log \frac{\exp(\text{sim}(i, j)/\tau)}{\sum_{k=1, k \neq i}^{2N} \exp(\text{sim}(i, k)/\tau)}, \quad (2)$$

where  $\tau$  denotes a temperature parameter, as in [32]. For each pair in every minibatch, the total loss is computed as follows:

$$L = \frac{1}{2N} \sum_{k=1}^N [\text{loss}(2k-1, 2k) + \text{loss}(2k, 2k-1)]. \quad (3)$$

As shown in Figure 1, we use ResNet [33] as the neural network encoder (F) to extract the visual features after data augmentation, and we use an MLP network (G) to map the output feature  $f$  to the space where contrastive loss is applied. The contrastive prediction task involves finding the other example  $j$  in examples  $S'(i \neq j)$  ( $n=2N-1$ ) for example  $i$  in each pair. In addition, we impose a restriction that every minibatch input must be a set of continuous frames



randomly selected from the video frames of a single subject. This restriction eliminates the possibility of finding existing macrofeatures between different faces during training and helps the model focus on the microfeatures of tics.

**2.4. Stage 2: Training the LSTM through Visual Features.** In Stage 2, we design a supervised learning framework based on the formerly trained neural network encoder (F). The LSTM network consists of a layer of LSTM with dropout and a fully connected layer with rectified linear unit (ReLU) activation. Specifically, we take a one-second-long pre-processed video clip as the input of this stage and randomly select  $k$  frames ( $k < 25$ ) to feed to F, which generates visual feature vectors. Every visual feature vector came from one frame of the input video clip and corresponded to one neuron of the LSTM layer. These visual feature vector sequences are then fed to the LSTM network to learn their temporal features and deeper spatial features to accomplish the classification task.

To alleviate the problem of imbalanced categories in our labeled data, we use the focal loss [34]  $L_f$ , which is defined as follows:

$$L_f = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \quad p_t = \begin{cases} p, & \text{if } y = 1, \\ 1 - p, & \text{otherwise,} \end{cases} \quad (4)$$

where  $y$  denotes a tic label,  $p$  is the output prediction of the LSTM network, and  $\alpha$  and  $\gamma$  are network hyperparameters.

### 3. Evaluation and Results

**3.1. Experimental Setup.** Chen et al. [31] showed that the simple operation of expanding the batch data volume can replace the more complex memory library training model [35, 36]. In this work, the two stages are trained separately. In Stage 1, we set the batch size to 512 and limited the single-input data to a randomly chosen person's continuous data due to the memory limitation of the training platform. For the neural network encoder, we used a modified ResNet18 model, with an input dimension of  $112 \times 112 \times 3$  and an output dimension of 512. The following MLP network consists of two layers: the first layer has 512 neurons and the second layer has 256 neurons. All the preprocessed unlabeled video datasets were randomly split into a training set (70%) and a validation set (30%) at the patient level. During training, we used the Adam optimizer [37] with an initial learning rate of  $3 \times 10^{-4}$  adjusted by setting the cosine annealing learning rate (LR) and a weight decay of  $10^{-6}$ . Considering the limitation of our dataset, we used the pretrained built-in weights of PyTorch [38]. The training procedure is stopped when the loss of validation set has no more drops within 10 epochs. Then the ResNet model is reused in Stage 2 to extract feature representations.

In Stage 2, each video clip generates a feature vector of clip-length  $\times$  512 through F. Here, clip-length is set as 16, which means that there are 16 frames randomly sampled from each video clip, matching the time-step setting in the LSTM network. The input size of the LSTM is 512 with a

drop rate of 0.8; the output size is 128; and the size of the fully connected layer is changed to match the number of classes in each classification task. Considering the limited amounts of labeled data in the study that can easily cause overfitting during training and validation, we adopted the leave-one-subject-out cross-validation scheme in Stage 2, which allows us to evaluate the differences between individual patients. The setting for the overall analysis of a single patient is in line with the real clinical scenario, which is beneficial for the subsequent comprehensive analysis. We assess the effectiveness of our proposed method by calculating the accuracy, precision, recall, F1-score, area under the receiver operating characteristic (ROC) curve (AUC\_ROC), area under the precision-recall curve (AUC\_PR), and a confusion matrix for each subject evaluation. In the two different classification tasks, we consider different cutoff conditions during the training procedure by observing the following indicators from the validation evaluation: (1) accuracy and (2) the F1-score of the tic category. In addition to the data used for experimental modeling, we also collected individual test video data beyond those used for training verification to verify the universality of the method.

The next subsections report the details of the results and provide discussions. Unfortunately, to the best of our knowledge, no public TS dataset for tic detection exists, which makes it difficult to compare the results of our method with other works. Instead, we applied another two kinds of supervised ConvNet architectures, convolutional 3D (C3D) [22], and temporal segment network (TSN) [39], for comparison.

**3.2. Classification Tasks.** C3D [22] is a simple yet effective model that uses 3D convolution kernels for spatiotemporal feature learning, and TSN [39] combines a sparse temporal sampling strategy and video-level supervision. They both achieved good performances for action recognition in videos when given limited training samples. As shown in Table 2, compared with the former two approaches C3D [22] and TSN [39], our method with the watch-accuracy strategy achieves the best performances, with an average accuracy of 94.87%, precision of 86.26%, and both recall and F1-scores of approximately 80%. These results illustrate the effectiveness of our proposed method for tic recognition on the multiclass classification task.

Using the classification results for an individual subject, we further examine the misclassified items. Taking Case 1 as an example, as shown in Figure 4, after checking the original data, we found that (a) in the false positive result where the label is normal but the prediction is mouth twitching, the mouth of the patient in this video clip does indeed twitch in the corners during a smile, indicating that the classification model has learned the features of the motion but cannot precisely differentiate between a mouth-twitching motion and a mouth-smiling motion when both are subtle; thus, it misclassifies the action. (b) In false negatives where the labels are eye tics while the prediction is normal, the patient in this video clip is indeed blinking, but it is difficult for ordinary people and for the model to determine whether the blink is



TABLE 2: Evaluations of the multiclass classification task.

	Accuracy	Precision	Recall	F1-score
C3D [22]	0.7252 ( $\pm 0.108$ )	0.7483 ( $\pm 0.047$ )	0.7023 ( $\pm 0.051$ )	0.7194 ( $\pm 0.032$ )
TSN [39]	0.8988 ( $\pm 0.117$ )	0.8354 ( $\pm 0.089$ )	0.7284 ( $\pm 0.054$ )	0.7600 ( $\pm 0.070$ )
Ours-acc <sup>1</sup>	<b>0.9487 (<math>\pm 0.0298</math>)**</b>	<b>0.8626 (<math>\pm 0.084</math>)**</b>	<b>0.7878 (<math>\pm 0.106</math>)*</b>	<b>0.7975 (<math>\pm 0.093</math>)*</b>
Ours-f1 <sup>2</sup>	0.9363 ( $\pm 0.0390$ )	0.7628 ( $\pm 0.209$ )	0.7362 ( $\pm 0.198$ )	0.7391 ( $\pm 0.198$ )

<sup>1</sup>Ours-acc means the proposed architecture with the watch-accuracy strategy. <sup>2</sup>Ours-f1 means the proposed architecture with the watch-F1 strategy. \*  $p$  value  $< 0.01$ ; \*\*  $p$  value  $< 0.001$ .

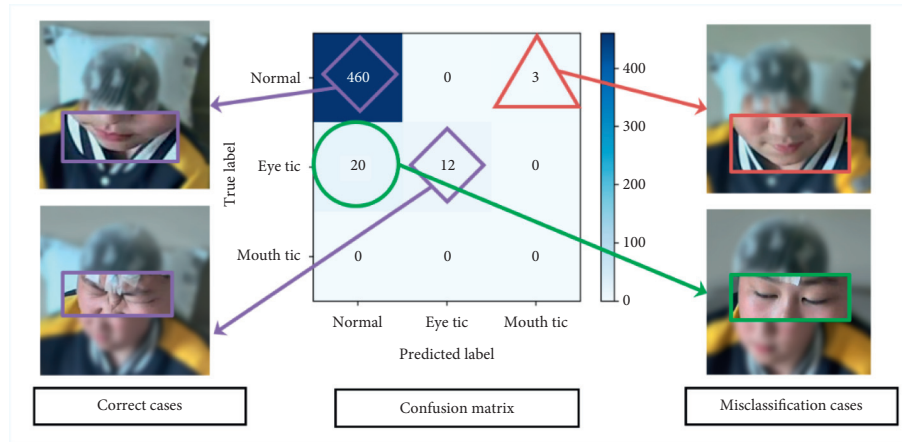


FIGURE 4: Evaluation result of one subject. The confusion matrix is shown in the middle; the correct detection cases from the multiclass classification task are shown on the left; and the misclassification cases are shown on the right. For the sake of patient privacy, the images used in the cases were blurred.

normal or a twitch blink. This may be due to model misunderstanding, but a small possibility of labeling error also exists. These two situations reflect either possible misclassifications or misdiagnosis in the real situation; however, it should be noted that the identification unit in this study is at the level of a video clip, whereas the identification unit in the clinic is the complete subject over time. A few misidentifications from the clip may not be completely reflected at the level of subject recognition. Therefore, quantitative methods and full quantification of the video data for the standard duration at the subject level will be considered in future research. This will allow individual subject evaluations to be made and improve the model's application prospects for clinical auxiliary diagnosis.

The evaluation results of the binary classification task are shown in Table 3. Compared with the multiclass classification task, the indicator results are slightly lower in Table 2, and its watch-F1 strategy performs better. After comparing the datasets of the two tasks, we find that the multiclass data use two types of tics with more discriminative characteristics, which provides a data quality advantage. In contrast, the positive samples in the binary classification task cover all the tic categories that appear in the TS dataset. Despite these shortcomings, our model still achieves good recognition performance and will offer substantial clinical value after further optimization.

**3.3. Further Evaluation.** To verify the visual representation performance of the unsupervised model in Stage 1, we visualized the final layer features of the neural network

encoder using the gradient-weighted class activation mapping (Grad-CAM) [40] method. As shown in Figure 5, the feature attention areas are shown as heatmap colors, and the attention areas are consistent with the corresponding tic positions, indicating that the unsupervised model has effectively learned the visual features used in the follow-up training.

To verify the validity of the proposed method and the possibility of subsequent integration with scales such as MRVRS [11], we compared the differences between the model's output and the clinician's result. This comparison test was based on the labeled dataset using the leave-one-subject-out test. We used two items based on the MRVRS and modified it within our data condition and one item for time comparison. The number of tic areas came from the annotations performed by clinicians and the tic categories of the model output. The tic frequency was calculated as the number of tic signs divided by the total length of the video used for every patient. Then, a  $t$ -test was performed on each of the items. The time for evaluation for clinicians was recorded between the start and the end for each video evaluation, and for our model it was calculated as the sum of the time taken for the whole process of our architecture, including preprocessing, model calculation, and postprocessing, among which preprocessing is the most time-consuming process. The subitem clinician review refers to the time taken for the clinician's checking process on the results of our models, which is divided into two categories:  $< 5$  min (0–5 min) and  $< 10$  min (5–10 min). The results are listed as Table 4. The  $p$  values of the two scale-related items are greater than 0.05, which shows no significant difference between the two groups of results. The  $p$

TABLE 3: Evaluations of binary classification task.

	Accuracy	AUC_ROC	AUC_PR	Precision	Recall	F1-score
Ours-acc	0.8890 ( $\pm 0.0458$ )	0.7532 ( $\pm 0.080$ )	0.7035 ( $\pm 0.138$ )	0.8057 ( $\pm 0.103$ )	0.7532 ( $\pm 0.103$ )	0.7634 ( $\pm 0.093$ )
Ours-f1	<b>0.9057 (<math>\pm 0.0479</math>)</b>	<b>0.7815 (<math>\pm 0.155</math>)</b>	<b>0.7669 (<math>\pm 0.187</math>)</b>	<b>0.8633 (<math>\pm 0.150</math>)</b>	<b>0.7707 (<math>\pm 0.296</math>)</b>	<b>0.7874 (<math>\pm 0.264</math>)</b>

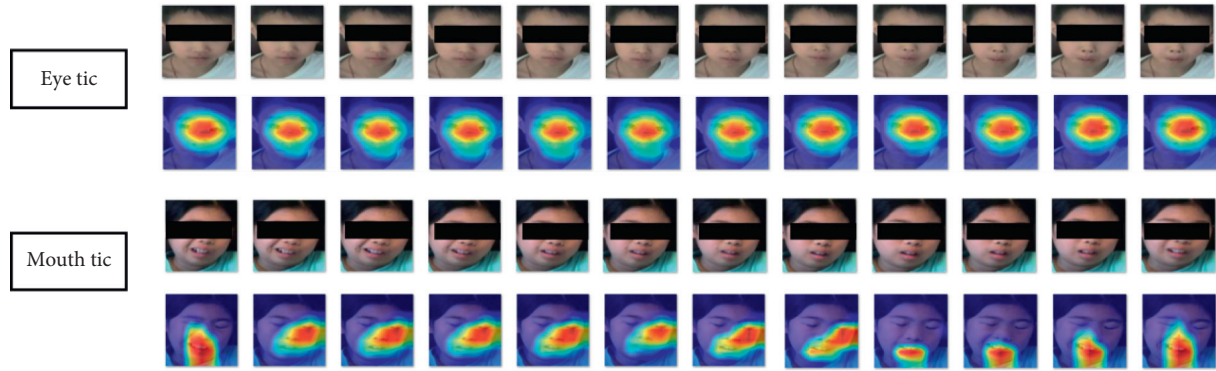


FIGURE 5: Visualization of two tic video clips of representation learning in Stage 1. The first row shows the original video clip frames; the second row shows the corresponding CAM image.

TABLE 4: Evaluations of some items of scales.

Test ID	Number of tic areas		Tic frequency (tics/min)		Time for evaluation (min)		
	Clinician	Our model	Clinician	Our model	Clinician	Our model	Clinician review
1	2	2	6	5	>40	<5	<5
2	2	2	6	7	>40	<5	<5
3	1	2	2	1	>30	<5	<5
4	1	1	40	37	>70	<5	<10
5	1	1	3	1	>30	<5	<5
6	1	1	9	12	>50	<5	<10
7	1	1	15	14	>60	<5	<5
8	1	1	0	0	>30	<5	<5
9	1	1	1	1	>30	<5	<5
10	1	1	11	8	>50	<5	<10
11	2	2	3	3	>30	<5	<5
12	1	0	0	0	>30	<5	<5
13	1	2	4	2	>40	<5	<5
<b>p value</b>	<b>0.7211</b>		<b>0.8666</b>		<b>&lt;0.0001</b>		-

value of the time comparison is less than 0.0001, showing that our model can save considerable time on tic detection, especially for videos with frequent tic events, which indicates great potential for clinical application.

Further evaluation for independent testing was additionally conducted on new data to identify tic events, including 1 new patient video and 4 non-TS patient videos. The binary classification task was adopted for this experiment. In the non-TS patient video testing, as shown in Table 5, the recognition accuracy for every non-TS patient was above 90%; the highest accuracy exceeded 98%. As discussed in the preceding subsection, the evaluation scores are computed at the video-clip level. If that were to be upgraded to the subject level in a clinical application, this level of individual evaluation would be acceptable. For the patient data, we must take recall into account; that is, the tic detection accuracy reaches 72.69%. Although we have only

TABLE 5: Non-TS patient evaluation.

No.	Accuracy	Number of clips
1	0.9701	67
2	0.9531	192
3	0.9016	193
4	0.9890	91

one testing video for this initial study, while these results lack statistical significance, they can still indicate that this approach has optimistic application prospects.

#### 4. Discussion

Tourette syndrome is a highly individualized neurological disease whose expression changes over time. In the process of long-term observation, diagnostic evaluation, and

management of the patient, the ability to continuously monitor and record tic events is the key to obtaining a patient-specific understanding of the disease. While reviewing and evaluating these monitoring data is a highly time- and cost-intensive process for doctors, the use of computer-assisted detection of tic movements can save time and cost, empower doctors to optimize and adjust medication responses, and help establish a good evaluation and management process for patients. Our work is the first application of a deep learning for video-based assessment of Tourette syndrome.

As the above experimental evaluations show, our video-based architecture possesses the ability to detect motor tic events in TS from videos acquired in a natural state. In the classification tasks, we detected two kinds of tics that occur most often in patients. Although the multiclass classification task involves limited motor tic categories in our dataset, it represents a unique result: to the best of our knowledge, no other similar research that has used surveillance video data for automatic tic recognition and classification exists. In the evaluation of subsequent results of the model, we defined some items that frequently appeared on the tic scales applied on these model outcomes and obtained consistent results with those from clinicians based on the MRVRS, which shows the ability to integrate observation-based scales or screening instruments for tics, although our current dataset limited a part of it. If we expand to audio data in the near future, it could be more comprehensive for developing an automatic rating scale of tics. For the binary classification task, it achieved good accuracy on video-clip-based recognition; however, it needs more video data for individual tests and other clinical data to support its outstanding performance in computer-aided diagnosis.

From our perspective, this work has application prospects from two main aspects: (a) automatic annotation of a video TS dataset. Because our classification task is based on small video clips, the task model can be used to prelabel the video and can then be checked by a doctor in a subsequent continuous data collection task, thereby reducing the doctor's labeling workload and accelerating the accumulation of labeled data. (b) Home-based health management applications: the extensive use of monitoring cameras makes it possible to extend this work to home-based health monitoring and management since acquiring video at home enhances the retrieval of objective tic expression [41]. In this case, object recognition and tracking, multiangle analysis, body tic detection, etc. all need to be considered and resolved. Furthermore, noise reduction and voice extraction are also significant for voice tic detection. A home-based tic surveillance system allows doctors and family members to better manage and provide more effective treatments for patients with tics who are undergoing long-term observation and treatment.

The inadequacy of labeled data is a clear limitation to future work and constitutes a weakness that we alleviate through unsupervised learning methods. We will continue

to try to ameliorate this limitation by integrating the few-shot learning method, which has performed well on many tasks with only small amounts of available training data [42, 43]. Moreover, this work can be applied and expanded to multicenter data analysis similar to [44, 45]; a larger research platform may result in additional interesting research works.

## 5. Conclusions

In this work, we introduce the first application of a deep learning method that combines unsupervised and supervised learning for video-based facial tic motion detection in TS patients. The developed model achieved good classification results on both multiclass and binary classification tasks; it can both detect and classify facial tic behaviors. This study effectively utilized large amounts of unlabeled data, which greatly reduced the labeling workload. A subsequent quantification of tic behavior has potential clinical application value for early identification and auxiliary diagnosis and evaluation of treatment effects. In the future, more video data will be collected and used to evaluate our scheme.

## Data Availability

The TS video data used to support the findings of this study are restricted by the Ethics Committee of the Second Affiliated Hospital of Zhejiang University School of Medicine to protect the patient privacy. The data are not publicly available due to ethical restrictions.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2018YFC0116901), the National Natural Science Foundation of China (Nos. 81771936 and 81801796), the Fundamental Research Funds for the Central Universities (No. 2021FZZX002-18), the Major Scientific Project of Zhejiang Lab (No. 2020ND8AD01), and the Youth Innovation Team Project of the College of Biomedical Engineering and Instrument Science, Zhejiang University.

## Supplementary Materials

Figure S1: Distribution diagram of the TS dataset, including the unlabeled dataset and the labeled dataset. The upper panel shows the frame distribution of every patient video, and the lower panel shows the time distribution of the videos in the TS dataset. (*Supplementary Materials*)

## References

- [1] M. M. Robertson, V. Eapen, H. S. Singer et al., "Gilles de la Tourette syndrome," *Nature Reviews Disease Primers*, vol. 3, no. 1, 2017.
- [2] B. S. Peterson and J. F. Leckman, "The temporal dynamics of tics in Gilles de la Tourette syndrome," *Biological Psychiatry*, vol. 44, no. 12, pp. 1337–1348, 1998.
- [3] L. Scahill, M. Specht, and C. Page, "The prevalence of tic disorders and clinical characteristics in children," *Journal of Obsessive-Compulsive and Related Disorders*, vol. 3, no. 4, pp. 394–400, 2014.
- [4] M. Bernabei, G. Andreoni, M. O. Mendez Garcia et al., "Automatic detection of tic activity in the Tourette Syndrome," in *Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 422–425, Québec, Canada, July 2010.
- [5] D. Martino, T. M. Pringsheim, A. E. Cavanna et al., "Systematic review of severity scales and screening instruments for tics: critique and recommendations," *Movement Disorders*, vol. 32, no. 3, pp. 467–473, 2017.
- [6] F. M. J. Pfister, T. T. Um, D. C. Pichler et al., "High-resolution motor state detection in Parkinson's disease using convolutional neural networks," *Scientific Reports*, vol. 10, no. 1, 2020.
- [7] D. Ahmedt-Aristizabal, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio, and C. Fookes, "Understanding patients' behavior: vision-based analysis of seizure disorders," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2583–2591, 2019.
- [8] B. Soran, L. Lowes, and K. M. Steele, "Evaluation of infants with spinal muscular atrophy type-I using convolutional neural networks," in *Proceedings of the Lecture Notes in Computer Science in Computer Vision – ECCV 2016 Workshops*, pp. 495–507, Amsterdam, The Netherlands, October 2016.
- [9] Y. Sun, D. Kommers, W. Wang et al., "Automatic and continuous discomfort detection for premature infants in a NICU using video-based motion analysis," in *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5995–5999, Berlin, Germany, July 2019.
- [10] J. F. Leckman, M. A. Riddle, M. T. Hardin et al., "The Yale Global Tic Severity Scale: initial testing of a clinician-rated scale of tic severity," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 28, no. 4, pp. 566–573, 1989.
- [11] C. G. Goetz, E. J. Pappert, E. D. Louis, R. Raman, and S. Leurgans, "Advantages of a modified scoring method for the rush video-based tic rating scale," *Movement Disorders*, vol. 14, no. 3, pp. 502–506, 1999.
- [12] H. Rickards, "Tics and fits. The current status of Gilles de la Tourette syndrome and its relationship with epilepsy," *Seizure*, vol. 4, no. 4, pp. 259–266, 1995.
- [13] P. B. Chappell, M. T. Mcswiggan-Hardin, L. Scahill et al., "Videotape tic counts in the assessment of tourette's syndrome: stability, reliability, and validity," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 33, no. 3, pp. 386–393, 1994.
- [14] D. Servello, M. Porta, M. Sassi, A. Brambilla, and M. M. Robertson, "Deep brain stimulation in 18 patients with severe Gilles de la Tourette syndrome refractory to treatment: the surgery and stimulation," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 2, pp. 136–142, 2008.
- [15] O. Rose, A. Hartmann, Y. Worbe, J. M. Scharf, and K. J. Black, "Tourette syndrome research highlights from 2018," *F1000Research*, vol. 8, p. 988, 2019.
- [16] T. Pringsheim, M. S. Okun, K. Müller-Vahl et al., "Practice guideline recommendations summary: treatment of tics in people with Tourette syndrome and chronic tic disorders," *Neurology*, vol. 92, no. 19, pp. 896–906, 2019.
- [17] J. B. Shute, M. S. Okun, E. Opri et al., "Thalamocortical network activity enables chronic tic detection in humans with Tourette syndrome," *NeuroImage: Clinical*, vol. 12, pp. 165–172, 2016.
- [18] A. Barua, C. Dong, and X. Yang, "A deep learning approach for detecting tic disorder using wireless channel information," *Transactions on Emerging Telecommunications Technologies*, vol. 10, Article ID e3964, 2020.
- [19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," 2014, <http://arxiv.org/abs/1406.2199>.
- [20] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1933–1941, Las Vegas, NV, USA, June 2016.
- [21] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1003–1012, Honolulu, HI, July 2017.
- [22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," 2014, <http://arxiv.org/abs/1412.0767>.
- [23] A. Diba, "Temporal 3D ConvNets: new architecture and transfer learning for video classification," 2017, <http://arxiv.org/abs/1711.08200>.
- [24] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? a new model and the kinetics dataset," 2018, <http://arxiv.org/abs/1705.07750>.
- [25] Z. Liu, "TEINet: towards an efficient architecture for video recognition," 2019, <http://arxiv.org/abs/1911.09435>.
- [26] J. Deng, W. Dong, R. Socher et al., "ImageNet: a large-scale hierarchical image database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, Florida, June 2009.
- [27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
- [28] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: a dataset for recognising faces across pose and age," 2018, <http://arxiv.org/abs/1710.08092>.
- [29] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," 2019, <http://arxiv.org/abs/1906.00910>.
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105, Curran Associates, Inc., New York, NY, USA, 2012.
- [31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, <http://arxiv.org/abs/2002.05709>.
- [32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, <http://arxiv.org/abs/1503.02531>.



- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, <http://arxiv.org/abs/1512.03385>.
- [34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2018, <http://arxiv.org/abs/1708.02002>.
- [35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2020, <http://arxiv.org/abs/1911.05722>.
- [36] Z. Wu, Y. Xiong, S. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," 2018, <http://arxiv.org/abs/1805.01978>.
- [37] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2020, <http://arxiv.org/abs/1412.6980>.
- [38] "torchvision.models — torchvision master documentation." <https://pytorch.org/vision/stable/models.html>.
- [39] L. Wang, "Temporal segment networks: towards good practices for deep action recognition," 2016, <http://arxiv.org/abs/1608.00859>.
- [40] R. R. Selvaraju, M. Cogswell, A. Das et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.
- [41] C. G. Goetz, S. Leurgans, and T. A. Chmura, "Home alone: methods to maximize tic expression for objective videotape assessments in Gilles de la Tourette syndrome," *Movement Disorders*, vol. 16, no. 4, pp. 693–697, 2001.
- [42] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni, "Generalizing from a few examples: a survey on few-shot learning," 2020, <http://arxiv.org/abs/1904.05046>.
- [43] D. Chen, Y. Chen, Y. Li, F. Mao, Y. He, and H. Xue, "Self-supervised learning for few-shot image classification," 2020, <http://arxiv.org/abs/1911.06045>.
- [44] J. Wang, Q. Wang, J. Peng et al., "Multi-task diagnosis for autism spectrum disorders using multi-modality features: a multi-center study," *Human Brain Mapping*, vol. 38, no. 6, pp. 3081–3097, 2017.
- [45] W. Wu, Y. Zhang, J. Jiang et al., "An electroencephalographic signature predicts antidepressant response in major depression," *Nature Biotechnology*, vol. 38, no. 4, pp. 439–447, 2020.



## Research Article

# Fp<sup>roi</sup>-GAN with Fused Regional Features for the Synthesis of High-Quality Paired Medical Images

Jiale Dong <sup>1,2</sup>, Caiwei Liu <sup>1,2</sup>, Panpan Man <sup>1,2</sup>, Guohua Zhao,<sup>1,2</sup> Yaping Wu <sup>3</sup>  
and Yusong Lin <sup>2,4,5</sup>

<sup>1</sup>School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China

<sup>2</sup>Collaborative Innovation Center for Internet Healthcare, Zhengzhou University, Zhengzhou 450052, China

<sup>3</sup>Department of Medical Imaging, Henan Provincial People's Hospital, Zhengzhou 450003, China

<sup>4</sup>School of Software, Zhengzhou University, Zhengzhou 450002, China

<sup>5</sup>Hanwei IoT Institute, Zhengzhou University, Zhengzhou 450002, China

Correspondence should be addressed to Yaping Wu; ypwu@ha.edu.cn and Yusong Lin; yslin@ha.edu.cn

Received 31 December 2020; Revised 20 February 2021; Accepted 16 April 2021; Published 28 April 2021

Academic Editor: Dong Nie

Copyright © 2021 Jiale Dong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The use of medical image synthesis with generative adversarial networks (GAN) is effective for expanding medical samples. The structural consistency between the synthesized and actual image is a key indicator of the quality of the synthesized image, and the region of interest (ROI) of the synthesized image is related to its usability, and these parameters are the two key issues in image synthesis. In this paper, the fusion-ROI patch GAN (Fp<sup>roi</sup>-GAN) model was constructed by incorporating a priori regional feature based on the two-stage cycle consistency mechanism of cycleGAN. This model has improved the tissue contrast of ROI and achieved the pairwise synthesis of high-quality medical images and their corresponding ROIs. The quantitative evaluation results in two publicly available datasets, INbreast and BRATS 2017, show that the synthesized ROI images have a DICE coefficient of  $0.981 \pm 0.11$  and a Hausdorff distance of  $4.21 \pm 2.84$  relative to the original images. The classification experimental results show that the synthesized images can effectively assist in the training of machine learning models, improve the generalization performance of prediction models, and improve the classification accuracy by 4% and sensitivity by 5.3% compared with the cycleGAN method. Hence, the paired medical images synthesized using Fp<sup>roi</sup>-GAN have high quality and structural consistency with real medical images.

## 1. Introduction

Medical imaging is a clinically important noninvasive diagnostic method; imaging specialists can diagnose breast cancer or precancer through mammography images [1]. With the development of deep learning technology, medical image synthesis [2, 3], classification [4], and segmentation [5] based on deep learning have become topical issues in medical research. Deep neural networks usually require a large number of training samples, and the size of medical image data is usually small because of the high collection cost, thus limiting the application of deep learning models for medical images [6]. Generative adversarial networks [7] usually learn feature mappings

from source modality to target modality by constructing generators and discriminators that can be used to synthesize medical images and thus expand training samples [8, 9]. However, the gradient disappearance, pattern collapse, and structural consistency problems between real and synthetic images in the current GAN research process seriously affect the quality of synthetic images [3]. In addition, the region of interest (ROI) of medical images is a key factor in aiding imaging research and is often used in training medical image segmentation tasks. However, we found that the synthesis of the ROI has rarely been studied [10, 11]. Thus, in the present study, we focused on the synthesis of high-quality medical images and their ROI images.

Nie et al. [12] were the first to propose a generative adversarial model using a fully convolutional neural network as a generator that implements the conversion between MRI and CT images of brain tumor images. The 3D-based fully convolutional neural network proposed in this paper well solves the problem of discontinuity across slices in 2D neural networks, and the method improves the quality of the generated images by calculating the gradient difference of the images as a loss function. The experimental results show that the method proposed in this paper can effectively predict CT images from MRI images, which is an early research and exploration of generative adversarial networks in the field of medical image synthesis. Guibas et al. [13] propose a novel pipeline model based on generative adversarial networks for the current medical images that are not easily accessible. The model proposed in this paper consists of Stage-I GAN and Stage-II GAN, which enables the generation of higher quality images by enhancing the learning of mask image features of images. In addition, John et al. created an online synthetic medical image database called SynthMed, while again demonstrating the feasibility of GAN-based synthesis of medical images. In addition, Chartsias et al. [8] proposed a multi-input, multi-output fully convolutional neural network for MRI synthesis, which embeds all input modalities into a shared potential space and converts the shared features into target output modalities by learning the potential space mapping through a decoder. Although this method can achieve multimodal output, the generated images are adulterated with redundant information. Wolterink et al. [9] used cycleGAN to learn the mapping of source modality to target modality through adversarial loss, resulting in synthetic CT images that are similar to the real CT images. Considering the lack of direct constraints between the real CT images and the synthesized CT images, this approach still cannot guarantee the structural consistency between the synthesized and the input images. Kang et al. [14] proposed a conditional GAN to improve model estimation and quantitatively evaluate the resulting images, but this approach resulted in uneven quality across domains of the synthesized images. Huang et al. [15] synthesized glioma images by using the WEENIE model, which uses a priori information instead of noise as input to the model, but the consistency of the synthesized images with the real images needs to be improved.

In the study of GAN-based generative models, the structural consistency between the real and synthetic images usually affects the quality of the synthetic images [3]. To improve the structural inconsistency between the real and synthesized image during image synthesis and synthesize the ROI of the image, we proposed a new method for synthesizing paired medical images based on cycleGAN. The method incorporates regional a priori features on the basis of cycleGAN two-stage cycle consistency to achieve high-quality medical images and their ROI synthesis. In the medical image synthesis process, the first stage model implements feature mapping from the medical image domain to the ROI domain and targets the learning contrast features of ROI and non-ROI tissues. The second stage network reduces the ROI domain to the medical image domain to

synthesize medical images. By contrast, in the synthesis process of ROI, the input ROI image is first reduced to a medical image, and then a high-quality ROI image is synthesized based on the regional contrast of the medical image. The two-stage synthesis process is implemented through the cycle consistency function of cycleGAN [16]. In this paper, we validated the quality of the synthesized images by using two publicly available datasets, where the benign data of the INbreast dataset has no corresponding ROI images. Then, we quantitatively analyzed the synthesis results from various metrics only. The results show that our proposed method effectively improves the structural consistency between the synthesized and real image, and the quality of the synthesized image is better than several recent popular models. In addition, we have verified that the images synthesized in this paper can improve the classification performance of the prediction model in the brain glioma classification experiment. The experimental results demonstrate that the method in this paper can effectively generate high-quality paired medical images, which will bring new solutions for medical disease research where it is difficult to obtain data.

The contribution of this work is summarized as follows.

- (1) We proposed a new synthesis method for the synthesis of paired medical images on the cycle consistency mechanism of cycleGAN and called it  $Fp^{roi}$ -GAN
- (2) To improve the quality of the synthesized images, this paper assists the generative model to learn ROI and non-ROI organizational features by supplementing a priori regional features
- (3)  $Fp^{roi}$ -GAN proved its effectiveness on two experimental datasets, and the experimental results show that our method can effectively improve the structural consistency of synthesized images with real images and outperform many popular image synthesis methods

## 2. Materials and Methods

**2.1. Dataset.** INbreast [17, 18] contains 303 normal (no mass) mammograms and 107 pairs of mammograms, including mass data and corresponding ROI images. Considering that training requires paired data, only 107 pairs of images containing masses were finally selected as the experimental data and then preprocessed. The mammograms had a resolution of  $3,328 \times 4,084$  pixels or  $2,560 \times 3,328$  pixels, and the images were stored in dicom format. We first cropped the original images according to the provided lesion areas, and the cropped images to  $256 \times 256$  were converted into PNG format, as shown in Figure 1(a). The processed paired data were divided into training and test sets in a ratio of 7:3, and the image intensity was linearly normalized to [0,1] by using maximum normalization. Subsequently, the influence of data irregularity on the experimental results was eliminated, and the network was accelerated to determine the optimal solution.

The BRATS 2017 [19, 20] dataset contains 285 medical images and their corresponding ROI images from four

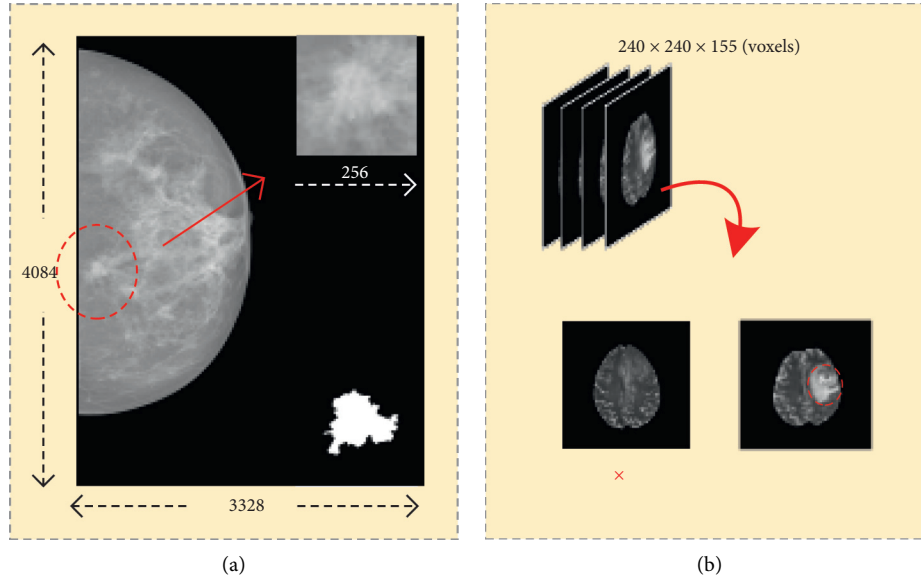


FIGURE 1: (a) Cropping of the INbreast. (b) Slices containing tumor regions were extracted from the 3D images of glioma; × indicates that images that do not contain tumor domains were excluded.

sequences, T1-weighted (T1), T1-weighted and contrast-enhanced (T1ce), T2-weighted (T2), and FLAIR, including 210 high-grade gliomas (HGG) and 75 low-grade gliomas (LGG), with image sizes of  $240 \times 240 \times 155$  voxels. T2 sequences were selected as the experimental data, and the 90th and 100th layer slices of the HGG images (the middle layer contains more brain image information relative to the edge of the voxel images) and the corresponding ROI slices were extracted first. Similarly, the 90th, 95th, 100th, and 105th layer slices of LGG images and the corresponding ROI slices were extracted, and each of the 272 pairs of HGG and LGG images were collected, as shown in Figure 1(b). Finally, all images were adjusted to  $256 \times 256$  pixels. The two small datasets, HGG and LGG, were normalized according to INbreast's partitioning and processing method.

**2.2. Methods.** To enable the network learn the contrast information of ROI and non-ROI tissues, we improved the cycleGAN model and proposed a pairwise image synthesis method that incorporates regional features. Figure 2 shows the flowchart of the model, where the input of the network is the medical image and its corresponding ROI. Before the network started training, the medical image matrix was first multiplied with its ROI image matrix to obtain the regional image containing only the tumor. Then, we designed a regional feature extraction block (RFB) to extract the semantic features of regional images and fuse the extracted regional features with medical images as the input of source domain  $X$  and ROI as the input of target domain  $Y$ . During network training, the model discriminates between ROI and non-ROI organizational features by learning the mapping of domain  $X$  to domain  $Y$ . The a priori regional features enhance the learning process and then reduces domain  $Y$  to domain  $X$  to synthesize medical images. ROI synthesis first reduces the mapping of domain  $Y$  to domain  $X$  and

synthesizes high-quality ROI images based on the mapping of domain  $X$  to domain  $Y$ . Figure 2(c) shows the synthesis of medical images, and the process can be represented as:  $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$  as shown in (i); similarly, the synthesis process of ROI can be represented as:  $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ , as shown in (ii). The model proposed in this paper is composed of two generators, namely,  $G$  and  $F$ , and two discriminators, namely,  $D_x$  and  $D_y$ .

**2.2.1. Regional Feature Extraction.** First, the medical image was multiplied with the ROI matrix to obtain the regional image, and the operation steps are shown in Figure 2(a). We designed the RFB for extracting high-level semantic features of the region image, and its structure is shown in Figure 3. The feature extraction block is a simple convolutional neural network consisting of two mirror fill layers, three convolutional layers, and one deconvolutional layer. In the network, the operational details of the three convolutional layers are zoomed into the right side of Figure 3, where the convolutional details include convolution, instance normalization, and activation operations. The feature map output after the RFB is fused with the medical image as the input of domain  $X$ .

**2.2.2. Network Architecture.** The  $Fp^{roi}$ -GAN model consists of two generators and two discriminators, where the structures of the generators  $G$  and  $F$  are shown in Figure 4. The generator consists of four convolutional layers, two fusion layers, and two deconvolutional layers, and the operation details of each convolutional layer include convolution, instance normalization, and activation operations. To extract each pixel in the fused image, the generator first performs a  $3 \times 3$  mirror fill of the image, and the feature map size is filled from  $256 \times 256$  pixels to  $262 \times 262$  pixels after filling. After three convolution processes, a 128-dimensional

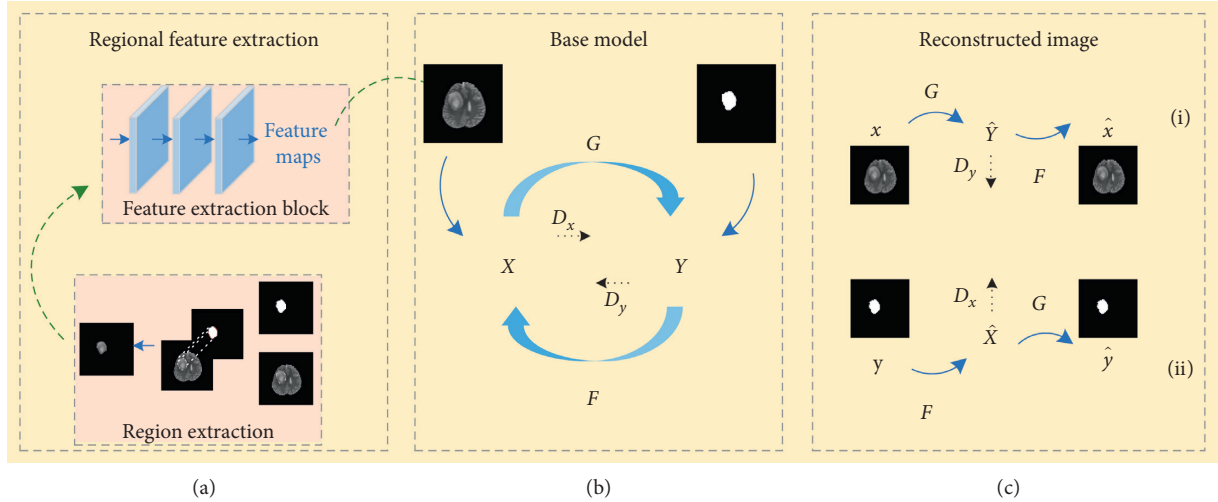


FIGURE 2: (a) Regional feature extraction method. (b) The base model is a like-cycleGAN model consisting of two generators and two discriminators. (c) Synthesis of paired images. (i) Process synthesis of medical images. (ii) Process synthesis of ROI images.

$64 \times 64$  feature map is obtained. The convolution aims to downsample the image and extracts its structural features, where the details of the three convolution layer operations are zoomed into the corresponding color boxes on both sides. In addition, we added two fusion layers to the generator to preserve the low-level image information. Finally, two deconvolution layers restore the image to its initial size and complete the image synthesis.

The inputs of the discriminator  $D_x$  include real and synthetic medical images, while the inputs of the discriminator  $D_y$  include synthetic and real ROI images. The discriminator consists of four convolutional layers, flatten layer, dense layer, and sigmoid activation layer. The

convolved image is flattened by the flatten layer, and the dense layer reduces the features to one dimension. Finally, the sigmoid function determines whether the image is synthetic or real, and the details of the discriminator layers are depicted in Figure 5. The discriminator is executed immediately after the output of the generator.

**2.2.3. Training Loss.** The loss functions used in the synthesis of the images include the traditional adversarial [7] and cycle consistency loss [16]. The model uses adversarial loss as the mapping function. The mapping function  $G: X \rightarrow Y$  and its discriminator  $D_Y$  are expressed in (1) as follows:

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim P_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim P_{\text{data}}(x)}[\log(1 - D_Y(G(x)))] \quad (1)$$

where  $G(x)$  generates an image similar to the  $Y$  domain, and  $D_Y$  distinguishes between the synthesized sample and the real sample. In this process,  $G(x)$  aims to distinguish between information from ROI and non-ROI tissue, resulting in subsequent  $F(G(x))$  restoration process.  $G$  aims to minimize this objective against an adversary  $D_Y$  that tries to maximize it, in which  $\min_G \max_{D_Y} \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y)$ . Similarly, in the restoration process of  $F(G(x))$ , a similar mapping function  $F: Y \rightarrow X$  learns the mapping from the ROI image to the medical image, in which  $\min_F \max_{D_X} \mathcal{L}_{\text{GAN}}(F, D_X, Y, X)$ , where  $D_X$  represents its discriminator.

Traditional adversarial losses can only intermittently learn the mapping function from domain  $X$  to domain  $Y$  or vice versa. To constrain the consistency of the real image with the synthetic image, we used a cycle consistency loss function in the model to enhance the reduction process. In Figure 2(c),  $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$  constrains the synthesis process of the medical image, while  $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$  constrains the synthesis process of the ROI image. These two components constitute the cycle consistency loss, as shown in the following:

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim P_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim P_{\text{data}}(y)}[\|G(F(y)) - y\|_1] \quad (2)$$

**2.3. Evaluation Measures.** The peak signal-to-noise ratio (PSNR) [21], structural similarity (SSIM) [22], and

multiscale structural similarity (MS-SSIM) [23] were used for the quantitative evaluation of the synthesized medical

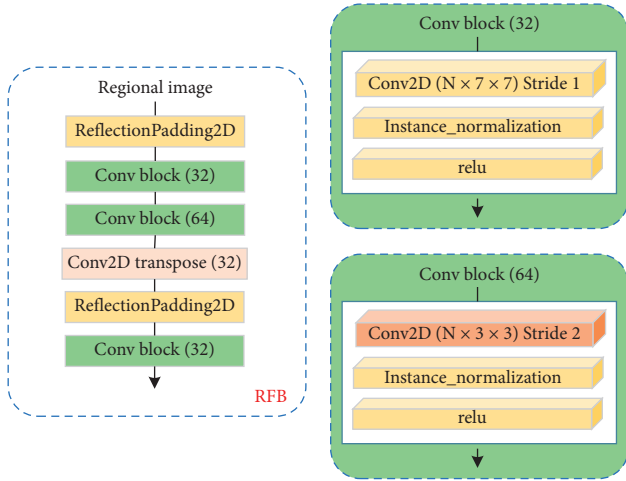


FIGURE 3: RFB architecture; the convolution process zoomed into the box on the right side corresponding to the dimension.

images. Dice coefficient and Hausdorff distance were used for the quantitative evaluation of the synthesized ROI images. Given the original input and synthetic images, the PSNR can be defined as follows:

$$\text{PSNR}(x, F(G(x))) = 10 \log_{10} \frac{\text{MAX}_{\text{range}}^2(x, F(G(x)))}{N_{\text{voxel}}^{-1} \|x - F(G(x))\|_2^2}, \quad (3)$$

$$\text{Hausdorff}(y, G(F(y))) = \max(\max_{y \in y} (\min(d(y, G(F(y))))), \max_{G(F(y)) \in G(F(y))} (\min(d(y, G(F(y))))), \quad (6)$$

where  $d$  represents the Euclidean distance.

### 3. Results and Discussion

Our network implementation was based on the PyTorch framework. All experiments were performed on a 12-core Intel Xeon 3.7 GHz CPU and GeForce RTX 2080 (8 GB) by using the Ubuntu 18.04 operating system. All figures were plotted on a computer with Windows 10 (8 GB) operating system. During the synthesis task, all models were trained for 300 epochs, where the trained models used the Adam optimizer [28] with default parameters, and the learning rate was set to 0.0002.

**3.1. Results of the INbreast Dataset.** This subsection provides a comparison of three commonly used synthesis models, namely, DCGAN [11], Pix2Pix [15], and cycleGAN [16]. Table 1 evaluates the whole and tumor domains of the synthesized images, and Table 2 compares the synthesis results of ROI images. Tables 1 and 2 compare the differences between Fp<sup>roi</sup>-GAN and the other methods using paired-samples  $T$ -tests [29], and the underline indicates a significant difference between Fp<sup>roi</sup>-GAN and the other methods at a significance level of 0.05. Based on the experimental results in Table 1, the Fp<sup>roi</sup>-GAN image synthesis method achieved

where  $\text{MAX}_{\text{range}}(x, F(G(x)))$  represents the maximum number of pixels for  $x$  and  $F(G(x))$  images, and  $N_{\text{voxel}}$  represents the total number of pixels for  $x$  or  $F(G(x))$ . The higher the PSNR value, the better the synthesis performance. SSIM was used to measure three metrics of image brightness, contrast, and structure, which can be expressed as follows:

$$\text{SSIM}(x, F(G(x))) = \frac{(2\mu_x\mu_{F(G(x))} + c_1)(2\sigma_{xF(G(x))} + c_2)}{(\mu_x^2 + \mu_{F(G(x))}^2 + c_1)(\sigma_x^2 + \sigma_{F(G(x))}^2 + c_2)}, \quad (4)$$

where  $\mu$  and  $\sigma^2$  denote the mean and variance of the image, respectively, and  $\sigma_{xF(G(x))}$  denotes the covariance of  $x$  and  $F(G(x))$ . The closer the SSIM is to 1, the higher the structural similarity is. The larger MS-SSIM values represent a better synthesis performance [24]. Dice coefficients [25, 26] are often used to represent the performance of the synthesized ROI image based on the ROI image  $y$  and the synthesized ROI images  $G(F(y))$  as follows:

$$\text{Dice}(y, G(F(y))) = \frac{2|y \cap G(F(y))|}{|y| + |G(F(y))|} \quad (5)$$

The Hausdorff distance [27], a complement to the Dice evaluation metric, can be expressed as follows:

the highest results for the three evaluation metrics, whereas the DCGAN synthesized image results were the lowest. Based on the quantitative analysis results of the whole image domain, the Fp<sup>roi</sup>-GAN values were 0.832, 0.053, and 0.016 higher than those of the cycleGAN method in the three evaluation metrics of PSNR, SSIM, and MS-SSIM, respectively, and 1.813, 0.113, and 0.056 higher than the DCGAN, respectively. In the tumor domain, the Fp<sup>roi</sup>-GAN values were 3.657, 0.085, and 0.042 higher than those of the cycleGAN method in the three evaluation metrics of PSNR, SSIM, and MS-SSIM, respectively, and 4.911, 0.095, and 0.052 higher than the DCGAN method, respectively. Fp<sup>roi</sup>-GAN method was significantly improved relative to other synthesis methods in Table 1. Based on the experimental results in Table 2, Fp<sup>roi</sup>-GAN obtained the highest DICE coefficient, which is 0.154 higher than DCGAN, and the lowest evaluated value in Hausdorff Distance, which is 3.10 lower than DCGAN. Figure 6 shows the visual performance of the four synthesis methods, and Fp<sup>roi</sup>-GAN performs closer to the original image in some detail positions.

**3.2. Results of the BraTS 2017 Dataset.** This subsection provides comparison with three commonly used synthetic models, such as DCGAN [11], Pix2Pix [15], and cycleGAN



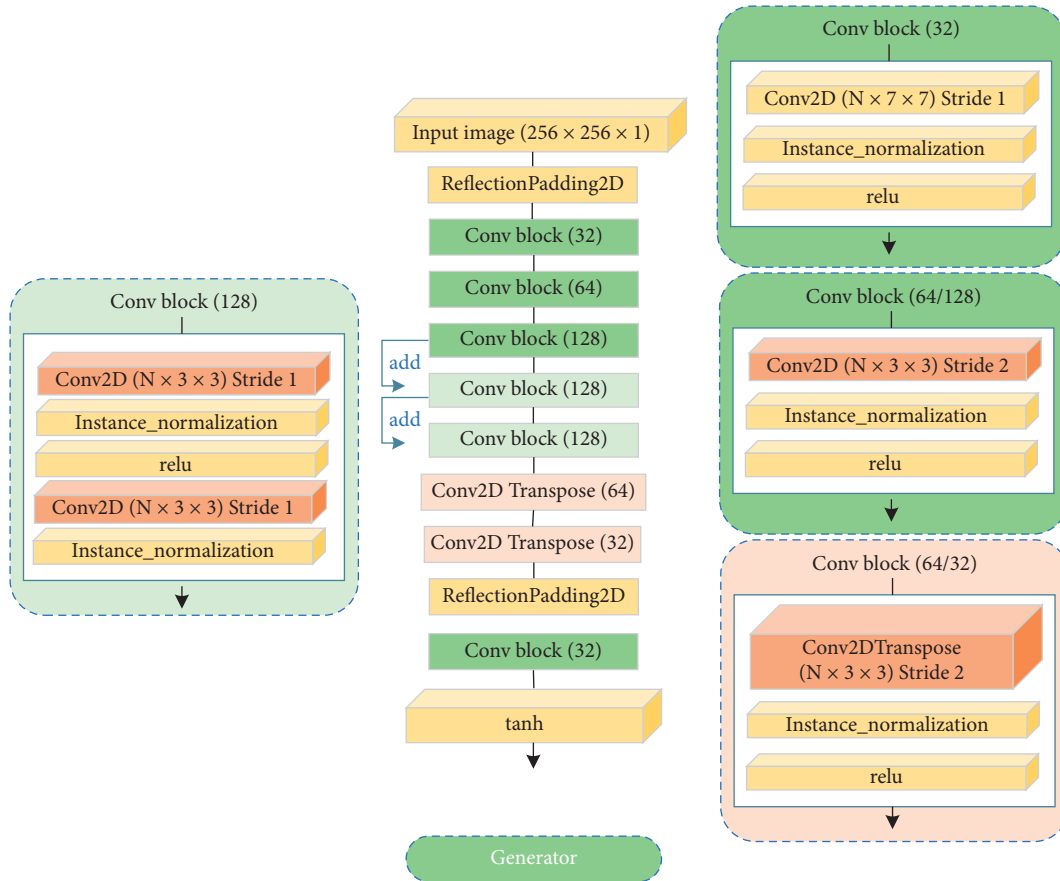


FIGURE 4: Generator architecture; the convolution details of the generator are zoomed into the boxes on both sides.

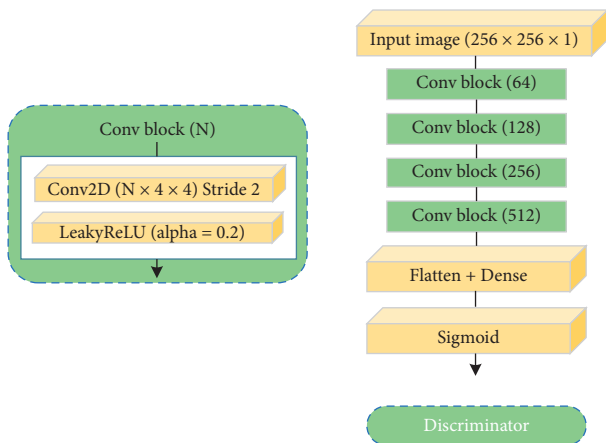


FIGURE 5: Discriminator architecture; Conv2D and LeakyReLU layers were applied to all Conv blocks.

[16]. Tables 3 and 4 compare the differences between  $Fp^{roi}$ -GAN and other methods by using paired-sample  $t$ -test [29], and the underline indicates that  $Fp^{roi}$ -GAN is statistically significantly different from other methods at a significance level of 0.05. Based on the experimental results in Table 3, the quantitative analysis results of  $Fp^{roi}$ -GAN in the HGG data for the whole image domain are higher than those of cycleGAN in PSNR, SSIM, and MS-SSIM by 0.604, 0.002,

and 0.003, respectively, and by 9.135, 0.104, and 0.097, compared with DCGAN, respectively. In the tumor domain, the  $Fp^{roi}$ -GAN values were higher than cycleGAN in PSNR, SSIM, and MS-SSIM by 6.236, 0.02, and 0.023, respectively, and 12.349, 0.094, and 0.083 higher than DCGAN method, respectively. The quantitative analysis results of  $Fp^{roi}$ -GAN in LGG data in the whole image domain are 1.999, 0.006, and 0.008 higher than cycleGAN in the three evaluation metrics of PSNR, SSIM, and MS-SSIM, respectively, and 6.951, 0.069, and 0.066 higher than DCGAN, respectively. In the tumor domain, the  $Fp^{roi}$ -GAN values were 11.248, 0.004, and 0.007 higher than cycleGAN and 14.631, 0.105, and 0.079 higher than DCGAN.

Based on the experimental results in Table 4,  $Fp^{roi}$ -GAN in HGG data achieved the highest DICE coefficient, which is 0.128 higher than DCGAN, and the lowest evaluated value in Hausdorff distance, which is 3.44 lower than DCGAN. The DICE coefficient of  $Fp^{roi}$ -GAN in LGG data part was higher than DCGAN by 0.101, and the Hausdorff distance was lower than DCGAN by 3.75. Figure 7 shows the visual results of the four synthesis methods, in which the synthesis results of the tumor domain, as well as the results of the non-tumor domain, are compared, as shown in the medical images of LGG. III shows the results of the synthesized paired images in ITK-SNAP [30], and the results show that  $Fp^{roi}$ -GAN method has less noise points than the other synthesis

TABLE 1: Quantitative evaluation of the INbreast dataset (mean  $\pm$  standard deviation). We compared the measurements of the different synthesis methods over the whole image domain and the tumor domain at a significance level of 0.05, and the underline indicates that  $Fp^{roi}$ -GAN is statistically significantly different from other methods.

Region	Methods	PSNR	SSIM	MS-SSIM
Whole image	DCGAN [11]	16.834 $\pm$ 3.28	0.769 $\pm$ 0.15	0.879 $\pm$ 0.21
	Pix2Pix [15]	17.398 $\pm$ 3.81	0.843 $\pm$ 0.13	0.923 $\pm$ 0.19
	cycleGAN [16]	17.815 $\pm$ 5.18	0.829 $\pm$ 0.17	0.919 $\pm$ 0.18
	<b><math>Fp^{roi}</math>-GAN</b>	<b>18.647 <math>\pm</math> 3.25</b>	<b>0.882 <math>\pm</math> 0.16</b>	<b>0.935 <math>\pm</math> 0.15</b>
Tumor region	DCGAN [11]	19.231 $\pm$ 7.43	0.872 $\pm$ 0.15	0.894 $\pm$ 0.23
	Pix2Pix [15]	21.811 $\pm$ 6.98	0.915 $\pm$ 0.11	0.902 $\pm$ 0.22
	cycleGAN [16]	20.485 $\pm$ 6.15	0.882 $\pm$ 0.07	0.904 $\pm$ 0.18
	<b><math>Fp^{roi}</math>-GAN</b>	<b>24.142 <math>\pm</math> 6.70</b>	<b>0.967 <math>\pm</math> 0.08</b>	<b>0.946 <math>\pm</math> 0.18</b>

TABLE 2: Results of the quantitative evaluation of the ROI images of the INbreast dataset (mean  $\pm$  standard deviation) with a significance level of 0.05; the underline indicates that the  $Fp^{roi}$ -GAN is statistically significantly different from other methods.

Methods	Dice coefficient	Hausdorff distance
DCGAN [11]	0.827 $\pm$ 0.25	7.31 $\pm$ 4.95
Pix2Pix [15]	0.945 $\pm$ 0.17	7.27 $\pm$ 4.18
cycleGAN [16]	0.952 $\pm$ 0.13	6.83 $\pm$ 3.38
<b><math>Fp^{roi}</math>-GAN</b>	<b>0.981 <math>\pm</math> 0.11</b>	<b>4.21 <math>\pm</math> 2.84</b>

methods. The results of the image distribution of the four synthesis methods are compared in Figure 8, where the histogram indicates the distribution of the image grayscale and the trend of the image grayscale. The  $Fp^{roi}$ -GAN method is always closer to the original image than the three other methods, both in terms of image distribution and trend of image grayscale.

In addition to the quantitative evaluation of the synthesized MR images, this paper supplements a glioma HGG and LGG classification experiment to verify the auxiliary effect of the synthesized MR images for the classification experiment. Considering that the INbreast dataset without mass data lacks corresponding ROI images, our synthesis method is not applicable, and the auxiliary effect on its dataset could not be verified in the classification. In the image synthesis experiments, the training set included 380 images, consisting of 190 HGG and LGG data, and the test set included 164 images, consisting of 82 HGG and LGG data. In the classification experiments, the data used for testing in the synthesis method were used as the training set with 164 images, the data used for training in the synthesis method were used as the test set with 380 images, and the data from each of the four groups synthesized images were added as comparison experiments, as shown in Table 5. Referring to article [31, 32] for the classification method, the first 500 features were extracted for each group of images by using the Resnet [33] network, followed by 30 features selected by the recursive feature elimination [34] with fivefold cross-validation, and the filtered features were classified using the kernel-based SVM algorithm [35]. The metrics used to assess the classification results include AUC, accuracy (Acc), sensitivity (Sen), and specificity (Spe), where AUC represents the area of the ROC curve and the other three metrics can be defined as (7)–(9):

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \quad (7)$$

$$Sen = \frac{TP}{TP + FN}, \quad (8)$$

$$Spe = \frac{TN}{FP + TN}, \quad (9)$$

where TP represents the number of samples, in which HGG was correctly predicted, TN represents the number of samples, in which LGG is correctly predicted, FN represents the number of samples, in which HGG is predicted as LGG, and FP represents the number of samples, in which LGG is predicted as HGG. The experimental results in Table 5 show that, by adding the images synthesized by our method for training the machine learning model, the prediction ability of the model was effectively improved, in which  $Fp^{roi}$ -GAN achieved the best results in the four metrics, and our method achieved a high classification sensitivity of 0.913. The ROC of the classification experiments is shown in Figure 9.

**3.3. Discussion.** Currently, most image synthesis methods are in single-input, single-output mode, and the synthesis of ROI images is rarely studied. Our work utilizes the cycleGAN's cyclic consistency mechanism to solve the problem of structural inconsistency between real and synthetic images and improves the contrast information between ROI and non-ROI domains by incorporating regional features a priori, resulting in the synthesis of high-quality medical images as well as the corresponding ROI images. To evaluate the quality of the synthesized images, we compared several currently popular synthesis methods, such as DCGAN, Pix2Pix, and cycleGAN, and evaluated the synthesis results in terms of the whole image domain of the images and the tumor domain. The results show that the  $Fp^{roi}$ -GAN method synthesized high-quality medical images on both datasets and achieved the best results in PSNR, SSIM, MS-SSIM, dice, and Hausdorff distance metrics. The poor quality of the DCGAN synthesized images may be due to the collapse of the model during training, and we found that the synthesized images of Pix2Pix and cycleGAN are not of high quality due to the low structural consistency of the model. In addition, the comparison results from the whole and tumor domain of

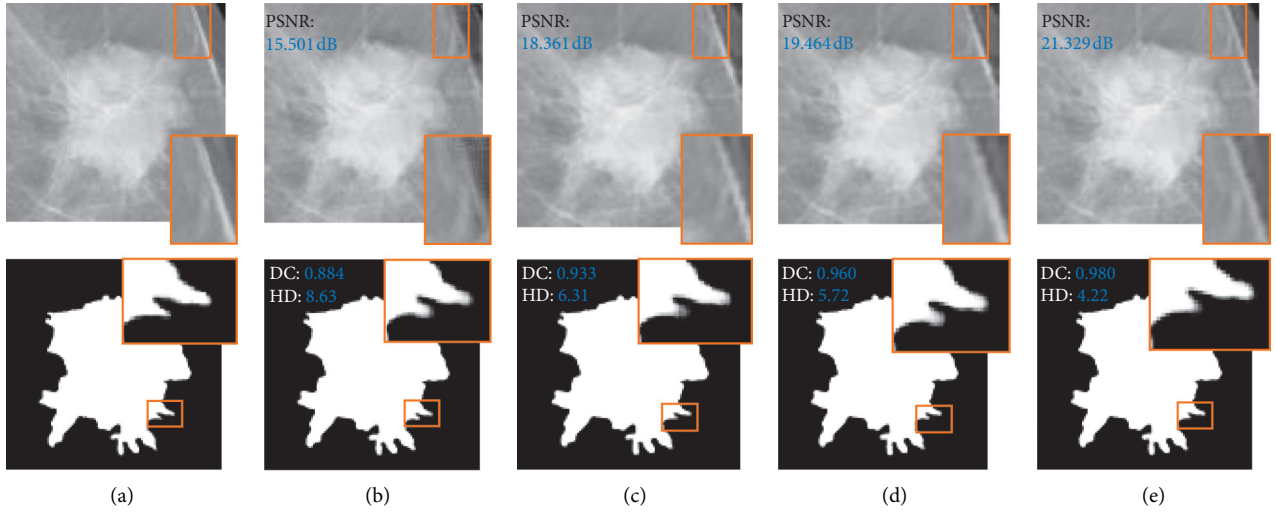


FIGURE 6: Comparison of  $Fp^{roi}$ -GAN with the three other synthesis methods on the INbreast dataset. (a) Input image. (b) DCGAN. (c) Pix2Pix. (d) cycleGAN. (e)  $Fp^{roi}$ -GAN.

TABLE 3: Results of the quantitative evaluation of the BRATS 2017 dataset (mean  $\pm$  standard deviation), where we compare the measurements of the different synthesis methods over the whole image domain and the tumor domain at a significance level of 0.05, and the underline indicates that  $Fp^{roi}$ -GAN is statistically significantly different from the other methods.

Data	Region	Methods	PSNR	SSIM	MS-SSIM
HGG	Whole image	DCGAN [11]	25.749 $\pm$ 3.49	0.882 $\pm$ 0.04	0.890 $\pm$ 0.05
		Pix2Pix [15]	28.938 $\pm$ 4.68	0.952 $\pm$ 0.03	0.956 $\pm$ 0.05
		cycleGAN [16]	34.280 $\pm$ 4.85	0.984 $\pm$ 0.02	0.984 $\pm$ 0.05
		<b><math>Fp^{roi}</math>-GAN</b>	<b>34.884 <math>\pm</math> 5.18</b>	<b>0.986 <math>\pm</math> 0.02</b>	<b>0.987 <math>\pm</math> 0.04</b>
	Tumor region	DCGAN [11]	29.539 $\pm$ 5.05	0.903 $\pm$ 0.02	0.910 $\pm$ 0.05
		Pix2Pix [15]	33.031 $\pm$ 5.99	0.951 $\pm$ 0.02	0.952 $\pm$ 0.04
		<b><math>Fp^{roi}</math>-GAN</b>	<b>41.888 <math>\pm</math> 6.06</b>	<b>0.997 <math>\pm</math> 0.004</b>	<b>0.993 <math>\pm</math> 0.03</b>
LGG	Whole image	DCGAN [11]	23.093 $\pm$ 4.71	0.895 $\pm$ 0.11	0.908 $\pm$ 0.06
		Pix2Pix [15]	25.912 $\pm$ 4.95	0.933 $\pm$ 0.09	0.945 $\pm$ 0.07
		cycleGAN [16]	28.045 $\pm$ 4.47	0.958 $\pm$ 0.08	0.966 $\pm$ 0.03
		<b><math>Fp^{roi}</math>-GAN</b>	<b>30.044 <math>\pm</math> 4.21</b>	<b>0.964 <math>\pm</math> 0.08</b>	<b>0.974 <math>\pm</math> 0.03</b>
	Tumor region	DCGAN [11]	25.809 $\pm$ 4.39	0.892 $\pm$ 0.09	0.911 $\pm$ 0.07
		Pix2Pix [15]	30.228 $\pm$ 5.28	0.939 $\pm$ 0.08	0.948 $\pm$ 0.07
		<b><math>Fp^{roi}</math>-GAN</b>	<b>40.440 <math>\pm</math> 7.51</b>	<b>0.997 <math>\pm</math> 0.02</b>	<b>0.990 <math>\pm</math> 0.03</b>

TABLE 4: Results of the quantitative evaluation of the ROI images of the BRATS 2017 dataset (mean  $\pm$  standard deviation) with a significance level of 0.05; underline indicates that the  $Fp^{roi}$ -GAN is statistically significantly different from other methods.

Data	Methods	Dice coefficient	Hausdorff distance
HGG	DCGAN [11]	0.808 $\pm$ 0.29	8.36 $\pm$ 5.66
	Pix2Pix [15]	0.876 $\pm$ 0.23	7.54 $\pm$ 5.90
	cycleGAN [16]	0.931 $\pm$ 0.18	5.15 $\pm$ 3.03
	<b><math>Fp^{roi}</math>-GAN</b>	<b>0.936 <math>\pm</math> 0.18</b>	<b>4.92 <math>\pm</math> 3.22</b>
LGG	DCGAN [11]	0.889 $\pm$ 0.26	7.83 $\pm$ 4.84
	Pix2Pix [15]	0.947 $\pm$ 0.23	6.25 $\pm$ 3.12
	cycleGAN [16]	0.984 $\pm$ 0.21	4.66 $\pm$ 2.33
	<b><math>Fp^{roi}</math>-GAN</b>	<b>0.990 <math>\pm</math> 0.25</b>	<b>4.08 <math>\pm</math> 2.79</b>

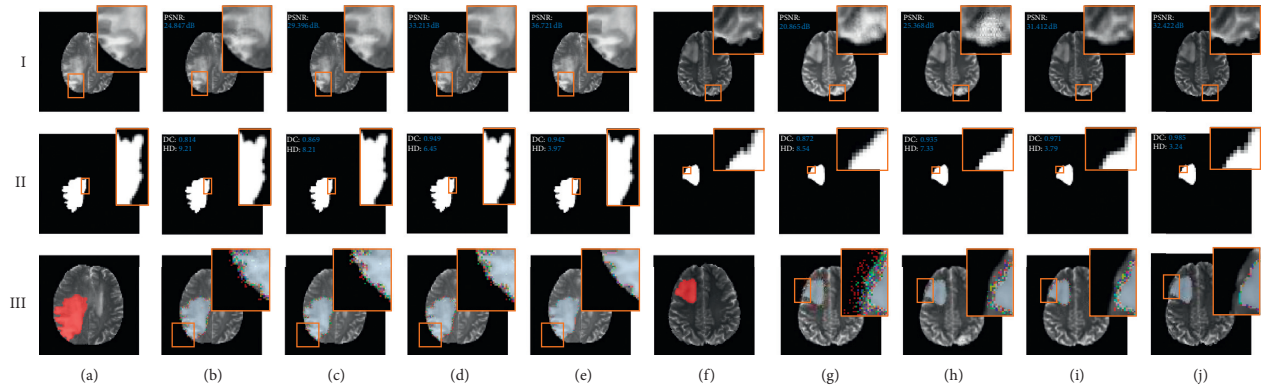


FIGURE 7: Comparison of  $Fp^{roi}$ -GAN with the three other synthesis methods on the BRATS 2017 dataset, where III is the visual performance in ITK-SNAP. (a) Input image (HGG). (b) DCGAN. (c) Pix2Pix. (d) cycleGAN. (e)  $Fp^{roi}$ -GAN. (f) Input image (LGG). (g) DCGAN. (h) Pix2Pix. (i) cycleGAN. (j)  $Fp^{roi}$ -GAN.

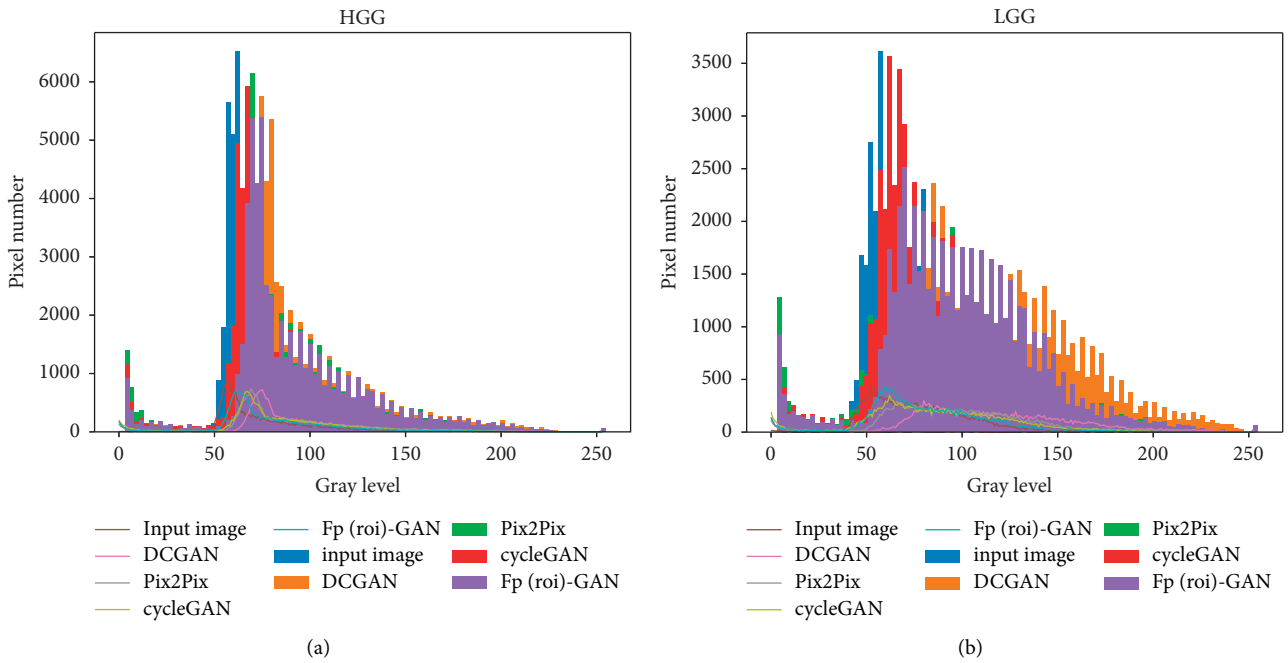


FIGURE 8: Image distribution results and grayscale trends of the four synthesis methods under HGG and LGG, where  $Fp^{roi}$ -GAN represents  $Fp^{roi}$ -GAN. (a) HGG. (b) LGG.

TABLE 5: Classification results.

Data	Methods	AUC	Acc	Sen	Spe
BRATS2017		0.872	0.789	0.823	0.778
BRATS2017 + DCGAN		0.881	0.803	0.720	0.831
BRATS2017 + Pix2Pix	Resnet + SVM	0.894	0.815	0.857	0.855
BRATS2017 + cycleGAN		0.928	0.855	0.910	0.843
BRATS2017 + $Fp^{roi}$ -GAN		<b>0.943</b>	<b>0.882</b>	<b>0.913</b>	<b>0.868</b>

the images showed that the tumor domain is more informative than the whole domain by incorporating regional features.

As shown in Figure 7, our synthesis method resulted in the least noise points in the medical image processing tool ITK-SNAP, but the images generated by DCGAN contain

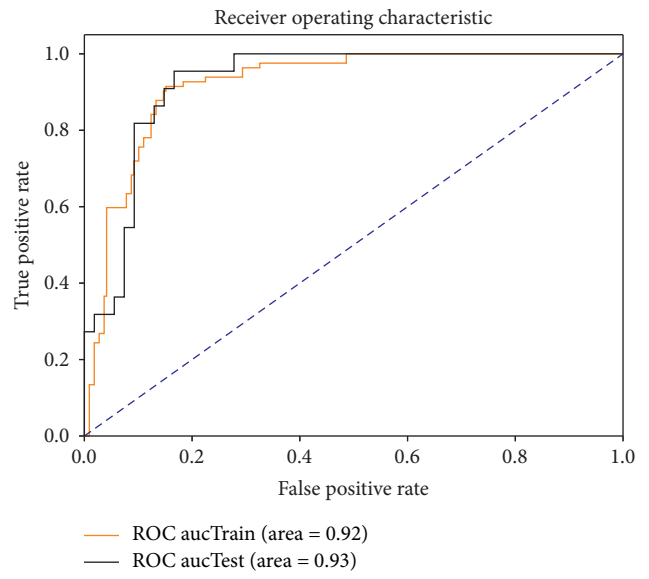
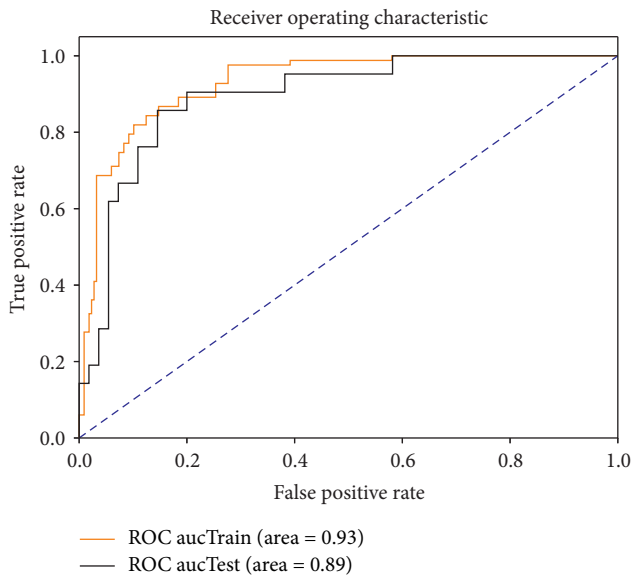
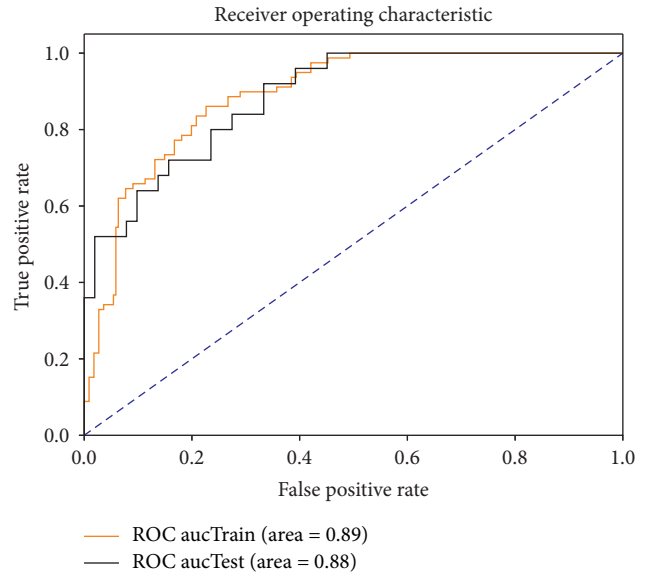
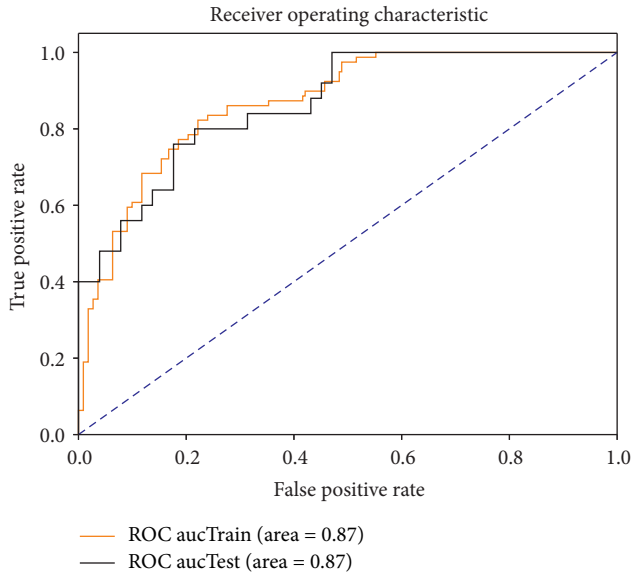


FIGURE 9: Continued.



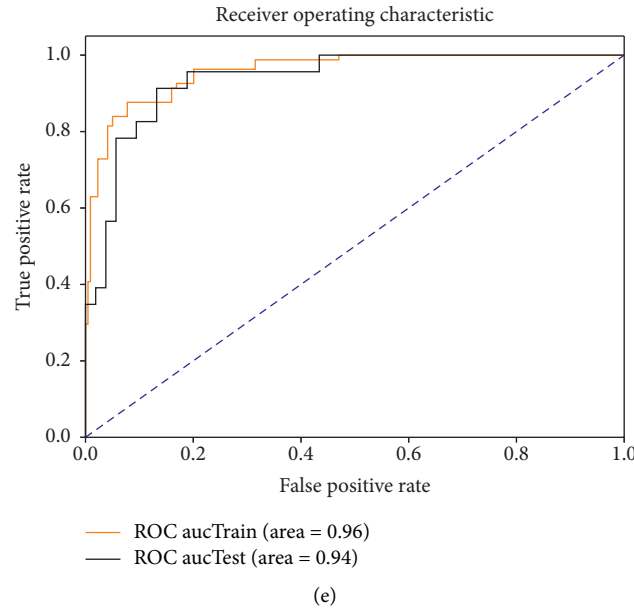


FIGURE 9: ROC plot of the classification experiment. (a) BRATS 2017. (b) BRATS 2017 + DCGAN. (c) BRATS 2017 + Pix2Pix. (d) BRATS 2017 + cycleGAN. (e) BRATS 2017 +  $Fp^{roi}$ -GAN.

more noise points. Based on the image distribution and grayscale change trend in Figure 8, the proposed method is closest to the distribution of the original image. From the previously mentioned experimental evaluation results, the sets of experiments show that the method proposed in this paper is more likely to be applied in the near future research of medical images. At last, in the BRATS 2017 classification experiments, we supplemented the synthesized data into the training set to effectively assist the training of the machine learning model and improve the classification effect of the model, in which the highest classification accuracy was achieved by adding the data synthesized by the  $Fp^{roi}$ -GAN method. Although many adversarial generation models have been proposed, the quality of the generated images has been an important goal for researchers to pay attention to, and in addition, whether the generated images can be used in recent studies is also a key concern for research. In this paper, our proposed method generates high-quality images and is validated in brain glioma classification experiments, which proximately illustrates the feasibility and superiority of our proposed generation method in the process of medical imaging research.

#### 4. Conclusions

GAN is widely studied in the field of medical imaging, including cross-modal synthesis, super-resolution reconstruction, and medical image denoising. In this paper, we proposed the  $Fp^{roi}$ -GAN method to synthesize paired medical images. Moreover, we validated the results of the synthesized images via quantitative analysis, image distribution comparison, and visual evaluation. In the BRATS experiment, we added a classification experiment to verify the effect of synthesized data on the classification experiment. The results show that the addition of synthetic images

effectively assisted the training of the machine learning model and improved the classification performance of the prediction model. Although this paper does not further validate the impact of the synthesized ROI images on the segmentation problem, the quantitative analysis indicated that our method has higher quantitative evaluation results than the other synthesis methods. In the future, we will further determine the effect of synthetic images on tasks, such as medical image classification and segmentation.

#### Data Availability

The datasets used in this paper are public dataset BRATS2017 and public dataset INbreast. BRATS2017 can be obtained through the following URL: <https://www.med.upenn.edu/sbia/brats2017/data.html>, and INbreast can be obtained through the following URL: <http://medicalresearch.inescporto.pt/breastresearch>.

#### Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

#### Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant no. 81772009 and Collaborative Innovation Major Project of Zhengzhou under grant no. 20XTZX06013.

#### References

- [1] Z. Zhou, Y. Wang, Y. Guo, Y. Qi, and J. Yu, "Image quality improvement of hand-held ultrasound devices with a two-stage

- generative adversarial network,” *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 298–311, 2020.
- [2] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, “CovidGAN: data augmentation using auxiliary classifier GAN for improved covid-19 detection,” *IEEE Access*, vol. 8, pp. 91916–91923, 2020.
  - [3] B. Yu, Y. Wang, L. Wang, D. Shen, and L. Zhou, “Medical image synthesis via deep learning,” *Advances in Experimental Medicine and Biology*, vol. 1213, pp. 23–44, 2020.
  - [4] C. Galván Casas, A. Català, G. Carretero Hernández et al., “Classification of the cutaneous manifestations of COVID-19: a rapid prospective nationwide consensus study in Spain with 375 cases,” *British Journal of Dermatology*, vol. 183, no. 1, pp. 71–77, 2020.
  - [5] N. Wang, S. H. Ma, J. Y. Li, Y. P. Zhang, and L. F. Zhang, “Multistage attention network for image inpainting,” *Pattern Recognition*, vol. 106, p. 107448, 2020.
  - [6] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, and J. Paisley, “An adversarial learning approach to medical image synthesis for lesion detection,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2303–2314, 2020.
  - [7] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.
  - [8] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsafaris, “Multimodal MR synthesis via modality-invariant latent representation,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 803–814, 2018.
  - [9] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije et al., “Deep MR to CT synthesis using unpaired data,” in *Proceedings of the International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 14–23, Québec City, Canada, September 2017.
  - [10] X. Xiao, J. Zhao, Y. Qiang et al., “Radiomics-guided GAN for segmentation of liver tumor without contrast agents,” in *Proceedings of the Medical Image Computing and Computer Assisted Intervention-Miccai*, pp. 237–245, Shenzhen, China, October 2019.
  - [11] T. Fujioka, M. Mori, K. Kubota et al., “Breast ultrasound image synthesis using deep convolutional generative adversarial networks,” *Diagnostics*, vol. 9, no. 4, p. 176, 2019.
  - [12] D. Nie, R. Trullo, J. Lian et al., “Medical image synthesis with context-aware generative adversarial networks,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 417–425, Québec City, Canada, September 2017.
  - [13] J. T. Guibas, T. S. Virdi, and P. S. Li, “Synthetic medical images from dual generative adversarial networks,” 2017, <http://arxiv.org/abs/1709.01872>.
  - [14] H. Kang, J.-S. Park, K. Cho, and D.-Y. Kang, “Visual and quantitative evaluation of amyloid brain PET image synthesis with generative adversarial network,” *Applied Sciences-Basel*, vol. 10, no. 7, p. 2628, 2020.
  - [15] Y. Huang, L. Shao, and A. F. Frangi, “Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding,” in *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5787–5796, Honolulu, HI, USA, July 2017.
  - [16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “UNPAIRED image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pp. 2242–2251, Venice, Italy, October 2017.
  - [17] I. Domingues, P. H. Abreu, and J. Santos, “BI-RADS classification of breast cancer: a new pre-processing pipeline for deep models training,” in *Proceedings of the 2018 25th IEEE International Conference on Image Processing*, pp. 1378–1382, Athens, Greece, October 2018.
  - [18] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, “INbreast: toward a full-field digital mammographic database,” *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.
  - [19] N. K. Batmanghelich, B. Taskar, and C. Davatzikos, “Generative-discriminative basis learning for medical imaging,” *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 51–69, 2012.
  - [20] B. H. Menze, A. Jakab, S. Bauer et al., “The multimodal brain tumor image segmentation benchmark (BRATS),” *Ieee Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.
  - [21] Q. Huynh-Thu and M. Ghanbari, “Scope of validity of PSNR in image/video quality assessment,” *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.
  - [22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
  - [23] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, vol. 2, pp. 1398–1402, Pacific Grove, CA, USA, November 2003.
  - [24] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, “Objective video quality assessment methods: a classification, review, and performance comparison,” *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.
  - [25] N. Ibtehaz and M. S. Rahman, “MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation,” *Neural Networks*, vol. 121, pp. 74–87, 2020.
  - [26] Y. Kong, J. Wu, G. Yang et al., “Iterative spatial fuzzy clustering for 3D brain magnetic resonance image supervoxel segmentation,” *Journal of Neuroscience Methods*, vol. 311, pp. 17–27, 2019.
  - [27] H. Berhane, M. Scott, M. Elbaz et al., “Fully automated 3D aortic segmentation of 4D flow MRI for hemodynamic analysis using deep learning,” *Magnetic Resonance in Medicine*, vol. 84, no. 4, pp. 2204–2218, 2020.
  - [28] D. Kingma and J. J. C. E. Ba, “Adam: a method for stochastic optimization,” 2014, <http://arxiv.org/abs/1412.6980v9>.
  - [29] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, “EaGANs: edge-aware generative adversarial networks for cross-modality MR image synthesis,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1750–1762, 2019.
  - [30] P. A. Yushkevich, Y. Gao, and G. Gerig, “ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images,” in *Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3342–3345, Orlando, FL, USA, August 2016.
  - [31] A. M. Ismael and A. Şengür, “Deep learning approaches for COVID-19 detection based on chest X-ray images,” *Expert Systems with Applications*, vol. 164, p. 114054, 2021.
  - [32] X. Yu, N. Zeng, S. Liu, and Y.-D. Zhang, “Utilization of DenseNet201 for diagnosis of breast abnormality,” *Machine Vision and Applications*, vol. 30, no. 7-8, pp. 1135–1144, 2019.
  - [33] L. Wen, X. Li, and L. Gao, “A transfer convolutional neural network for fault diagnosis based on ResNet-50,” *Neural*

*Computing and Applications*, vol. 32, no. 10, pp. 6111–6124, 2020.

- [34] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.
- [35] Z. Liu, Z. Li, J. Qu et al., “Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study,” *Clinical Cancer Research*, vol. 25, no. 12, pp. 3538–3547, 2019.

## Research Article

# MAGAN: Mask Attention Generative Adversarial Network for Liver Tumor CT Image Synthesis

Yang Liu,<sup>1</sup> Lu Meng ,<sup>2</sup> and Jianping Zhong<sup>2</sup>

<sup>1</sup>Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110000, China

<sup>2</sup>College of Information Science and Engineering, Northeastern University, Shenyang 110000, China

Correspondence should be addressed to Lu Meng; [menglul982@gmail.com](mailto:menglul982@gmail.com)

Received 8 December 2020; Revised 10 January 2021; Accepted 20 January 2021; Published 31 January 2021

Academic Editor: Jialin Peng

Copyright © 2021 Yang Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For deep learning, the size of the dataset greatly affects the final training effect. However, in the field of computer-aided diagnosis, medical image datasets are often limited and even scarce. We aim to synthesize medical images and enlarge the size of the medical image dataset. In the present study, we synthesized the liver CT images with a tumor based on the mask attention generative adversarial network (MAGAN). We masked the pixels of the liver tumor in the image as the attention map. And both the original image and attention map were loaded into the generator network to obtain the synthesized images. Then, the original images, the attention map, and the synthesized images were all loaded into the discriminator network to determine if the synthesized images were real or fake. Finally, we can use the generator network to synthesize liver CT images with a tumor. The experiments showed that our method outperformed the other state-of-the-art methods and can achieve a mean peak signal-to-noise ratio (PSNR) of 64.72 dB. All these results indicated that our method can synthesize liver CT images with a tumor and build a large medical image dataset, which may facilitate the progress of medical image analysis and computer-aided diagnosis. An earlier version of our study has been presented as a preprint in the following link: <https://www.researchsquare.com/article/rs-41685/v1>.

## 1. Introduction

Medical image analysis and processing is the core of computer-aided diagnosis, which has been greatly prompted by deep learning. And the training of deep learning can be extensively influenced by the size of the dataset; that is, the more datasets can be obtained, the better the performance the trained deep learning model can achieve. However, in the field of computer-aided diagnosis, the medical image is very limited and even scarce, due to the privacy of patients, the expense of medical image acquisition, and so on. Therefore, synthesized medical images can be seen as the only feasible way to solve this problem, and generative adversarial networks (GAN) [1, 2] provide us a powerful tool to realize it.

GAN was firstly proposed by Goodfellow and colleagues in 2014 and was widely used in various fields, such as image processing, natural language processing, and even medical image synthesis [3]. For skin lesion images, Baur and

colleagues synthesized the images of skin lesions with GAN [4], which enlarged the skin image dataset and improved the performance of lesion segmentation. For liver CT images, GAN was mainly used for expanding the dataset of the liver lesion [5] or image denoising [6], but the focus of GAN was only on the liver lesion, not on the whole liver CT images. For brain images [7], there are many image modules, such as CT images, magnetic resonance (MR) images, and positron emission tomography (PET), and different modules have different image acquisition methods and different influences on human brains. Dong Nie and colleagues used GAN to synthesize 7T images from 3T MR images [8] because 7T magnetic resonance (MR) images were very rare due to the expensive image acquisition costs and the side effects of high magnetic field strength. Moreover, some studies proposed to train a GAN to generate CT images from MR images to avoid the radiation from the CT image acquisition [9, 10]. For retinal images, the image resolutions were generally smaller than  $100 \times 100$ , and the image contents were only limited to

single color background and vessels. Based on the characteristics, some studies [11] used GAN to synthesize the whole retinal image to enlarge the retinal image dataset, but the method cannot be generalized to other medical image modules with bigger image resolution and more organs, such as liver CT image or brain MR image.

Above all, all these medical image synthesis methods can be categorized into three types: (1) transformation of different modules, such as from CT images to MR images, (2) transformation between the different parameter of image acquisition, such as from 3T MR images to 7T MR images, and (3) image synthesis of the small resolution, such as skin and retinal images. Although there were many existing methods, medical image synthesis is far from clinical applications, since there are still some shortcoming.

*1.1. Image Resolution.* Many current medical image synthesis methods can only synthesize images with low resolution, which were lower than  $128 \times 128$ . However, most of the medical images in the clinical application were high image resolution, such as  $512 \times 512$  CT images and  $512 \times 512$  MR images.

*1.2. Lesions or Tumors.* The current existing medical image synthesis methods cannot synthesize images with abnormalities, such as liver lesions and liver tumors. As we know, the size and variety of the training dataset are essential to the performance of deep learning methods. During the training of medical images' classification and analysis, it was essential to have both normal images and abnormal images to create an effective data set, but the medical images with abnormalities were relatively rare due to the hospital policy, patients' privacy, and so on. Therefore, synthesizing medical images with abnormalities can enlarge the dataset of deep learning methods and upgrade the performance.

To solve the shortcomings, we proposed a novel image synthesis model for normal liver CT images and liver CT images with tumors based on mask attention generative adversarial network (MAGAN). Using this model, we can build a liver CT image dataset consisting of thousands of synthesized  $512 \times 512$  slices; furthermore, it also can facilitate the progress of computer-aided diagnosis and the training of deep learning models.

The main contributions of our work are as follows: (1) we combined GAN with attention mechanism and proposed a novel MAGAN model and (2) we proposed an effective method of enlarging the existing medical image dataset.

## 2. Materials and Methods

In the present study, we synthesized liver CT images with tumors based on the mask attention generative adversarial network (MAGAN) model [12], whose framework is shown in Figure 1. Firstly, all the pixels of liver tumors in the original image were labeled by the white color and used as the attention map. According to the attention mechanism, liver tumors were the highlighted relevant features of the CT images, and the attention map was also the key part of the

success of the proposed algorithm. In the procedure of image synthesis, the liver tumor was the saliency map in the whole liver CT image, which meant that all the pixels of the liver tumors were masked by the attention map. The original image and the attention map were paired together and called "pairing A." Then, the original image and the attention map were loaded into the generator network to obtain a synthesized image, and the attention map and the synthesized image were paired together and called "pairing B." Next, pairing A and pairing B were both loaded into the discriminator network to determine if the synthesized image was real or fake. The generator network and the discriminator network were trained with adversarial learning so that both of them can become more and more powerful. After training, the generator network can fill the pixels of the attention map with similar gray values, texture, and shape of liver tumors, to synthesize liver CT images with tumors. More details of our model can be obtained from Sections 2.1~2.3.

*2.1. Attention Model.* All liver CT images used in our method were from a public liver CT dataset, Liver Tumor Segmentation (LiTS) [13, 14], which was from the MICCAI 2017 competition. In the LiTS dataset, the pixel distance was from 0.55 mm to 1.0 mm, the slice spacing was from 0.45 mm to 6.0 mm, and the image resolution was  $512 \times 512$ . LiTS consisted of 131 enhanced CT image sequences, and all the tumors in the liver CT images were manually labeled by radiologists. We aimed to synthesize liver CT images with tumors, and the synthesized materials were from two aspects, liver CT images from healthy controls and liver tumor CT images from patients. Moreover, the liver tumor was the most salient region for clinicians and was also the most difficult part of the whole synthesis procedure. Therefore, according to the tumor labels from the LiTS dataset, the image values of all the corresponding pixels in the tumors were changed to 4096, which meant "white color," and represented as an attention map in our model. Based on the attention mechanism, the original image and the attention map were transformed into feature maps  $A$  and  $B$  by using  $1 \times 1$  convolution, respectively, and then all these feature maps were concatenated by using matrix multiplication, shown in Figure 2:

$$S_{i,j} = A_i^T B_j. \quad (1)$$

Then, we performed softmax on the concatenated feature maps  $S_{i,j}$  to calculate the distribution of attention  $D_{i,j}$  on the  $i$ th position of the  $j$ th synthetic region:

$$D_{i,j} = \frac{\exp(s_{i,j})}{\sum_{i=1}^N \exp(s_{i,j})}. \quad (2)$$

Therefore, the liver tumor mask images were used as attention maps to efficiently find the liver tumors' internal and external characteristics of the images.



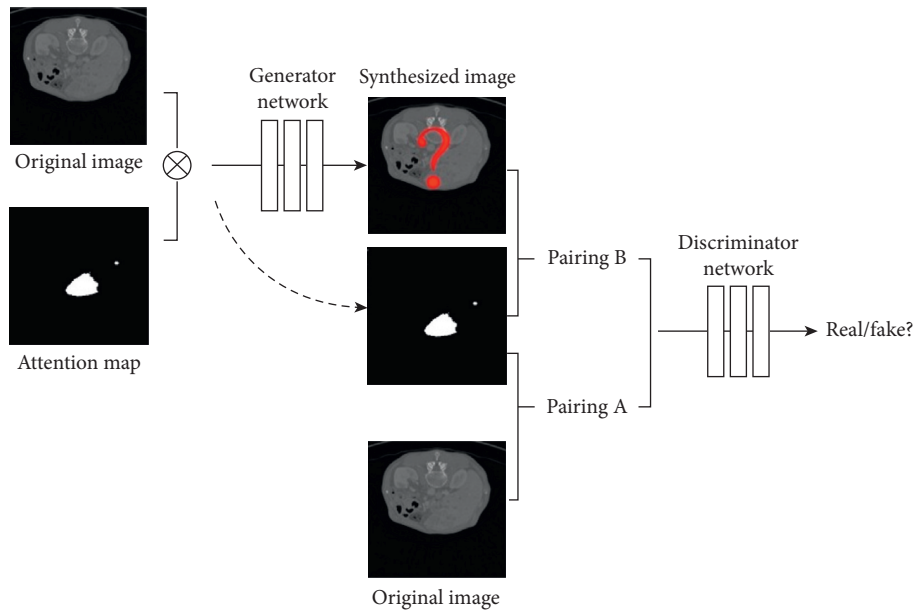


FIGURE 1: The framework of our model:  $\otimes$  represents matrix multiplication.

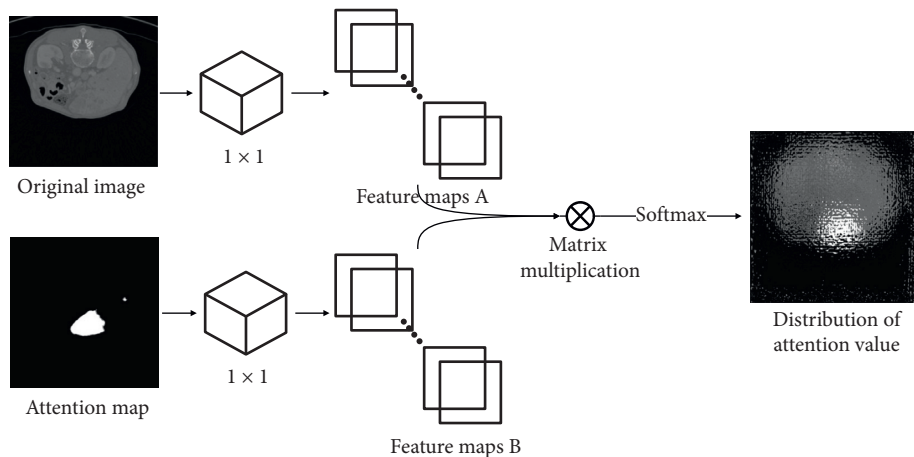


FIGURE 2: The framework of the attention model.

**2.2. Generator Network.** The structure of our generator network is shown in Figure 3, which consisted of two contracting paths and an expansive path, showing the U-shape architecture [15]. The input of these two contracting paths was the original image and attention map, respectively; both of them consisted of nine blocks, and each block was composed of the ReLu layer, convolutional layer, and batch normalization (BN) layer.

In the contracting path, the image resolution was reduced but the feature information was increased. To overcome the drawback of a regular convolution operator, whose receptive field was small, we used a dilated convolution operator [16] in the first four layers of the contracting path, so that we can capture image features from a larger scale. And we used a regular convolution operator in the other five layers of the contracting path because the sizes of the images were already smaller than  $32 \times 32$ , which cannot support a

dilated convolution operator. The feature maps from both of the two contracting paths were firstly loaded as input to the attention model, whose framework is shown in Figure 2, and then the distribution of attention value was transferred via residual connections. In the expansive path, the spatial information and the feature information were combined through a sequence of upconvolutions layer, BN layer, ReLu layer, and residual connections with high-resolution features from the attention model. Residual connections played important roles in MAGAN, which were used to bypass the nonlinear transformation, accelerate the training speed, and upgrade the performance of our model in the training of the deep CNN.

$512 \times 512$  original image and attention map were loaded as inputs into the generator network, and the image resolution was reduced by half while passing each block in the contracting path. After nine blocks in the contracting path,

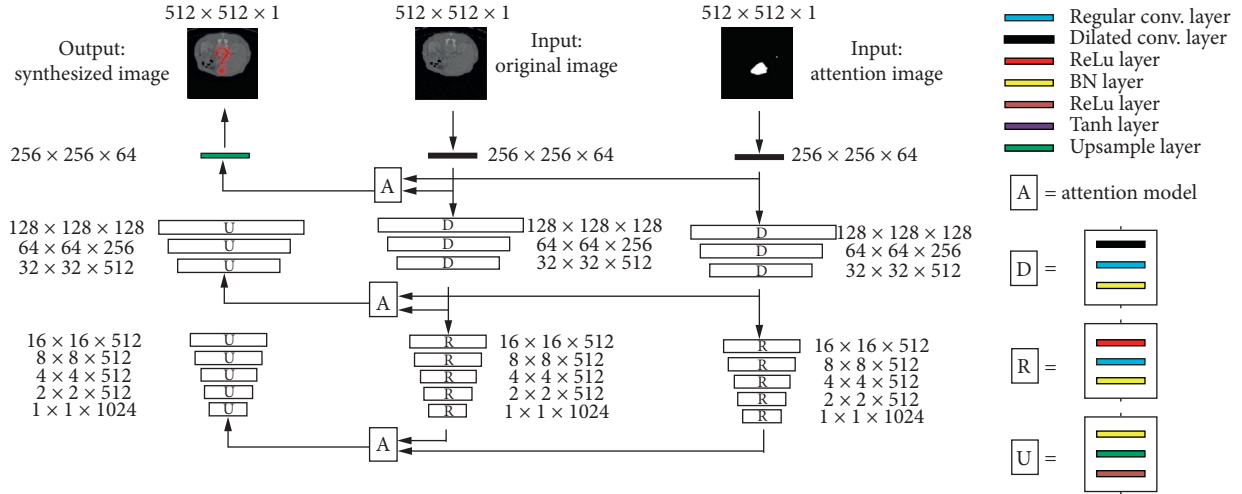


FIGURE 3: The framework of our generator network.

the input images became  $1 \times 1$  with 1024 feature maps. Then, these feature maps were upconvolved in the expansive path, and the size of the image increased one time while passing each block in the expansive path. After nine blocks in the expansive path, the image was restored as a  $512 \times 512$  resolution image. In the generator network, the whitened regions in the liver CT images can be transformed into tumor regions. The loss function of our generator network is shown as the following formula:

$$L_{adv}(G) = E_{v,r \sim p_{data}(v,r)} [\|r - G(v)\|_1], \quad (3)$$

where  $r$  denotes the real image,  $v$  denotes the concatenated image, and  $G(v)$  denotes the synthesized image calculated by the generator network.

**2.3. Discriminator Network.** The structure of our discriminator network is shown in Figure 4, which consisted of six

blocks, and each block was composed of a convolutional layer, ReLu layer, BN layer, or sigmoid layer.

The inputs of the discriminator network were two pairings, which were pairing A (original image, attention map) and pairing B (synthesized image, attention map). Inspired by PatchGAN [12], all the  $512 \times 512$  resolution images were divided into 900 patches, whose size was  $142 \times 142$ . After going through six blocks of the discriminator network, the sizes of output probabilities maps were  $30 \times 30$ , which indicated each pixel in the output probabilities maps corresponded to one patch of the input images. The mean value of all the pixels in the probabilities maps can be recognized as the result of the discriminator network.

The loss function of our discriminator network is shown as the following formula:

$$L_{adv}(D) = E_{v,r \sim p_{data}(v,r)} [\log D(v,r)_{real}] + E_{v \sim p_{data}(v)} [\log(1 - D(v, G(v,r))_{fake})], \quad (4)$$

where  $r$  denotes the real image,  $v$  denotes the attention map,  $G(v,r)$  denotes the synthesized image calculated by the generator network, and  $D(v,r)$  denotes the discrimination probability calculated by the discriminator network.

The total loss function of our GAN is shown as the following formula:

$$L = \arg \min_G \max_D \lambda_1 L_{adv}(G) + \lambda_2 L_{adv}(D), \quad (5)$$

where  $\lambda_1$  and  $\lambda_2$  are coefficients.

### 3. Results

In our experiments, we used LiTS as our image dataset of liver CT images with tumors, which consisted of only 131 sequences. The size of LiTS was not big enough for the

training of deep learning algorithms, such as liver tumor segmentation or classification. To enlarge the LiTS, we chose 4555 2D slices with tumors from 131 sequences of liver CT images. Then, all the images were normalized by using the following formula:

$$\text{value}_{\text{normalized}} = \frac{\text{value}_{\text{original}} - \text{mean}}{\text{std}}, \quad (6)$$

where  $\text{value}_{\text{original}}$  and  $\text{value}_{\text{normalized}}$  represent the original and normalized image pixels value, respectively. Mean indicate the mean value of image pixels, and std indicate the standard deviation of image pixels. Moreover, we specially cut the tumor regions from the liver CT images and built a liver tumor dataset; then, we augmented the tumor dataset by flipping, rotating, and scaling the original tumor region so that we can create a liver tumor dataset of 50000 slices from

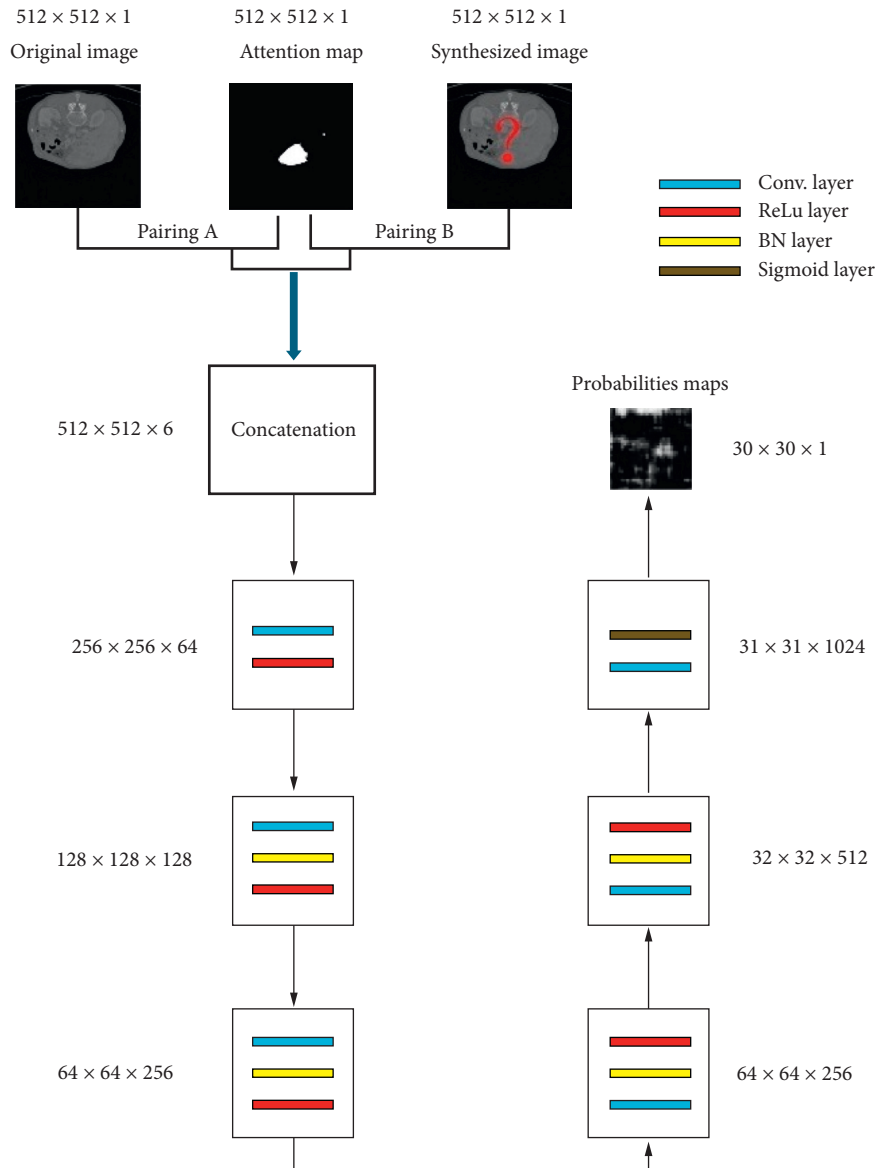


FIGURE 4: The framework of our discriminator network.

the original 4555 slices, which were used as the mask attention map in our method.

The hardware and software configuration of our experiments are shown in Table 1. The quantitative evaluation metric used in our experiments was the peak signal to noise ratio (PSNR). There were four sections in our experiments, including training of our model, quantitative comparison between our method and other state-of-the-art methods, Turing test for the synthesized images by radiologists, and the evaluation of the synthetic dataset for the medical image segmentation.

**3.1. Training of Our Model.** The configurations of hyperparameters in our model during the training are shown in Table 2. The proposed MAGAN network was implemented by Python 2.7 and TensorFlow 1.1 and trained on an NVIDIA GeForce GTX 1080 GPU using Adam optimizer

with a learning rate of 0.0002. It costs about ten hours for the whole procedure of the training.

As shown in Figures 5(a)–5(d), we can find that, as the number of iterations increased, the performance of the synthesized CT liver images became better and better. After the first iteration of training (in Figure 5(a)), the performance of the synthesized image from the generator network was terrible; for example, most pixels were black and the contour was blurring, intense chessboard effect. All these bad performances indicated that the training had just started, and more iterations were needed. After ten iterations (in Figure 5(b)), the whole image was more clear, the contour was more vivid, but the chessboard effect still existed. After one hundred iterations (in Figure 5(c)), the performance of the synthesized image was much better and closer to the real image, more details can be visualized, human organs were vivid, the chessboard effect was weaker but still existed, and whitened regions were not filled with

TABLE 1: Hardware and software configuration of our experiments.

Item	Configuration
Operating system	Ubuntu 16.04
GPU	NVIDIA GeForce GTX 1080
CPU	Intel Core i5-7500 @3.4 GHz
Software toolkit	Python 2.7; TensorFlow 1.1; MATLAB 2016b
Disk	500 GB
GPU memory	8 GB
System memory	16 GB

TABLE 2: Hyperparameters of our model.

Parameter	Value
Initial learning rate	0.0002
Adam momentum	0.5
$\lambda_1$ in formula (5)	100
$\lambda_2$ in formula (5)	1
Exponential decay	0.99
Batch_size	1
Epoch	10
Dropout	0.5
Frequency of saving loss value	100
Frequency of saving model	500

tumor texture. After one thousand iterations (in Figure 5(d)), the chessboard effect disappeared, all details of liver CT were restored, and it was hard to tell the differences between synthesized image and real image.

The loss function of the generator network, discriminator network, and total network during the training is shown in Figures 6–8, respectively, and we can conclude that the loss functions decreased as the number of iterations increased and became steady after about 10000 iterations, which indicated that our model performed well during the training.

Results of the synthesized image are shown in Figure 9: three liver tumor images with tumor masks were in the first row, which was used as inputs of our model, and we can obtain the synthesized images in the second row. We compared the synthesized images and the real images and calculated the differences between them. The color image of the differences is shown in the fourth row. All these results showed that our method can synthesize liver CT images with tumors, and the synthesized images were almost identical to the real images.

To test the impact of the dilated convolution operators in the MAGAN, we replaced the dilated convolution operators with the regular convolution operators in the contracting path of the generator network and quantitatively compared the PSNR of these two GAN networks. And we found that the network with regular convolution operators can provide a PSNR of 59.66, while the MAGAN with dilated convolution operators can provide a PSNR of 64.72, which indicated the effectiveness of the dilated convolution operators in our network.

To test the impact of the residual connections in the MAGAN, we removed the residual connections and quantitatively compared the PSNR of these two GAN

networks. And we found that the network without residual connections can provide a PSNR of 55.23, while the MAGAN with residual connections can provide a PSNR of 64.72, which indicated the effectiveness of the residual connections in our network. The running time of the proposed method was 0.087 seconds per frame.

Besides, we can also manually or automatically “add” tumor regions on the healthy liver CT images using our liver tumor dataset of 50000 slices, to create a diseased liver CT image, shown in Figure 10. The healthy liver CT images were in the first row. In the second row, manually change the pixel values of two regions to white color, which meant that these two regions were the selected tumor regions. Using our method, the results of the synthesized images are shown in the third row. All these results showed that our method can intelligently create liver CT images with tumors based on the healthy liver CT images, and the synthesized diseased images were almost identical to the real ones.

**3.2. Quantitative Comparison.** In this section, we quantitatively compared our method with other seven state-of-the-art medical synthesis methods using the same dataset as ours: (1) atlas-based method [17]; (2) sparse representation (SR) based method; (3) structured random forest with ACM (SRF+) [18]; (4) manipulable object synthesis (MOS) [19]; (5) deep convolutional adversarial networks (DCAN) method [8]; (6) multiconditional GAN(MC-GAN) [20]; and (7) mask embedding in conditional GAN (ME-cGAN) [21]. The first four methods were implemented by our group, and the source codes of DCAN, MOS, and ME-cGAN were downloaded from GitHub (<http://www.github.com/ginobilinie/medSynthesis>, [http://www.github.com/HYOJINPARK/MC\\_GAN](http://www.github.com/HYOJINPARK/MC_GAN), and [http://www.github.com/johnryh/Face\\_Embedding\\_GAN](http://www.github.com/johnryh/Face_Embedding_GAN)). The results of the quantitative comparison are shown in Table 3, which indicate that our method outperformed the other seven approaches and benefited from attention mechanism, dilated convolution operator, and residual connections.

**3.3. Turing Test.** To further verify the effectiveness of our method, we did the Turing test. Two experienced radiologists from Shengjing Hospital of China Medical University were asked to classify one hundred liver CT images into two types: real image or synthesized image. The radiologists were not aware of the answer to each image before the Turing test. The one hundred liver CT images consisted of fifty real CT images and fifty synthesized images. The results of the Turing test are shown in Table 4: radiologist number 1 made correct judgments for 74% real image slices and 64% synthesized image slices and radiologist number 2 made correct judgments for 84% real image slices and 12% synthesized image slices. The radiologists made correct judgments for most of the real images and may be psychologically influenced by the existence of a synthesized image, so they made some errors about the real images. Furthermore, the radiologists made difficult judgments for the synthesized images and cannot tell the obvious differences between the real images and the synthesized images. And according to radiologist #1, his

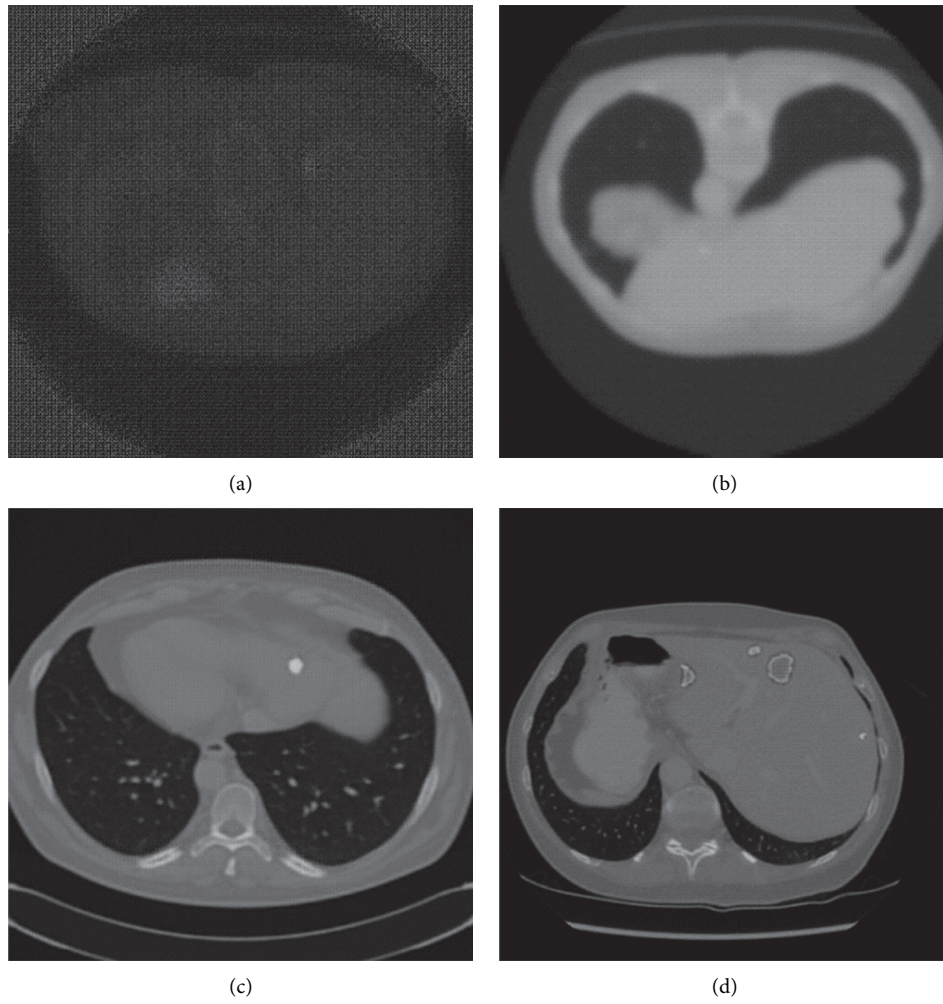


FIGURE 5: Synthesized image during the training of the proposed model: (a) after one iteration of training, (b) after ten iterations of training, (c) after one hundred iterations of training, (d) after one thousand iterations of training.

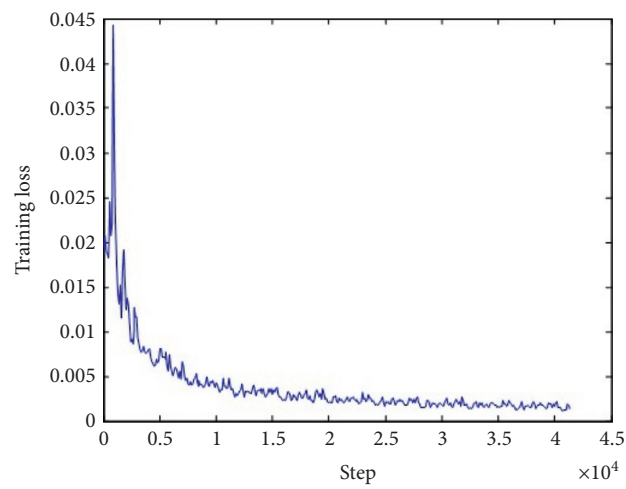


FIGURE 6: The loss function of the generator network during the training.



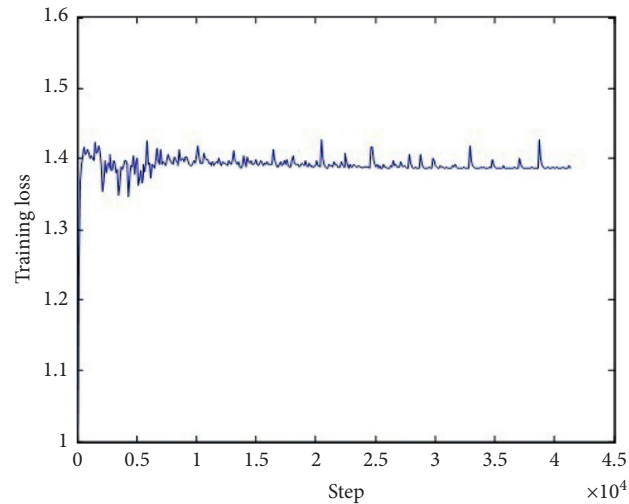


FIGURE 7: The loss function of the discriminator network during the training.

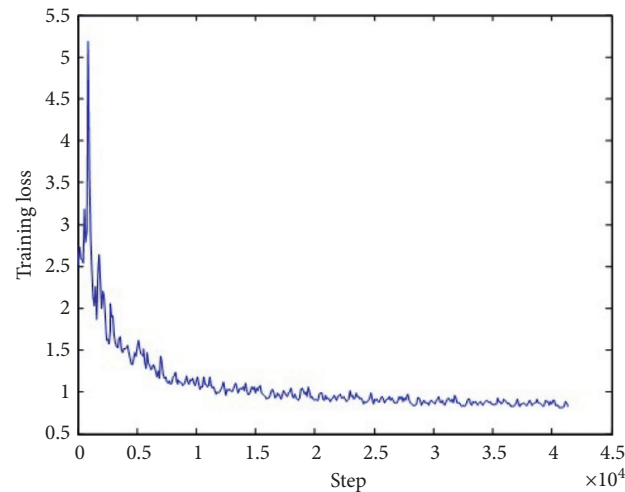


FIGURE 8: The loss function of the total network during the training.

most reliable evidence of telling the difference was the color of the tumor region was a little darker than the real ones, which was also the improvement we needed to do in the future. All these results of the Turing test indicated that our method can synthesize liver CT images with a tumor, which were almost identical to the real ones.

**3.4. Evaluation of Synthetic Dataset for Medical Image Segmentation.** To evaluate the effectiveness of the synthetic dataset in the training of deep learning models, we used a fully connected network (FCN) [15] to perform the tumor segmentation task in the liver CT images and trained the FCN model using the LiTS dataset (images from 131 subjects) and the new dataset obtained by our method (images from 131 real subjects and 865 synthetic subjects). And we used the Dice Index to quantitatively evaluate the performance of the segmentation results from the two trained FCN models. The FCN model trained by the LiTS dataset can provide a Dice value of 0.611 for the tumor segmentation,

and the FCN model trained by a new dataset can provide a Dice value of 0.658 for the tumor segmentation. The result indicated that the synthesized liver CT images obtained by the proposed method can effectively enlarge the original dataset, and as the number of images in the dataset increased, the performance of the training of the deep learning model can become better, which resulted in the higher Dice value for the liver tumor segmentation.

#### 4. Discussion

In the present study, we combined the attention mechanism and GAN model and proposed a novel CT image synthesis algorithm, which was MAGAN. As far as we know, the existing medical image synthesis methods mainly focused on the transformation of different modules or transformation between the different parameter of image acquisition, and our study was the first research of synthesizing the liver CT images with tumors in high resolution and enlarging the size of the medical image dataset.

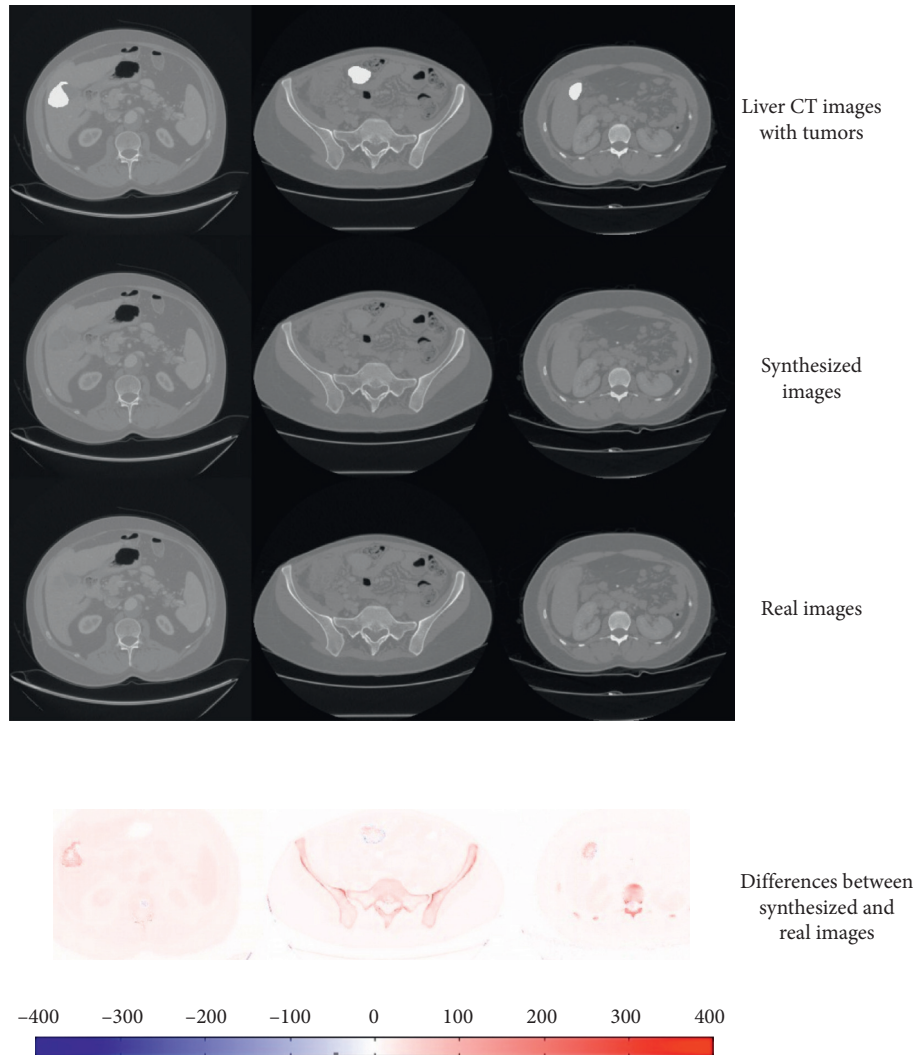


FIGURE 9: Results of the synthesized images and the comparison between the synthesized images and real images. The pixel values of the fourth rows are weak and low because the differences between the real images and synthesized images were very small.

Suppose that we had a dataset of chest CT images with lung nodules, whose size was one hundred. While we used this dataset for the training of deep learning, we may find that the trained model was not good enough due to the small size of the dataset. Under these circumstances, the proposed MAGAN can be used to synthesize thousands of chest CT images with lung nodules based on the original one hundred images. This kind of similar requirements from clinical researches and deep learning studies is very common. And the proposed method can meet the requirements.

From the quantitative comparison between the proposed method and the other seven state-of-the-art medical image synthesized methods, we can conclude that the proposed

method outperforms the others, and the main reasons were the attention map, which mainly focused on the regions of interest in the medical images, such as liver tumors or lung nodules.

During the Turing test, two experienced radiologists cannot clearly distinguish the synthesized liver CT images and the real liver CT images. We used the judgments of experts as the golden standard, and we may conclude that the synthesized liver CT images with tumors can be used as the real ones, and the size of the training dataset of medical images can be enlarged from one hundred to thousands. The bigger the medical image dataset is, the better the training performance can be.

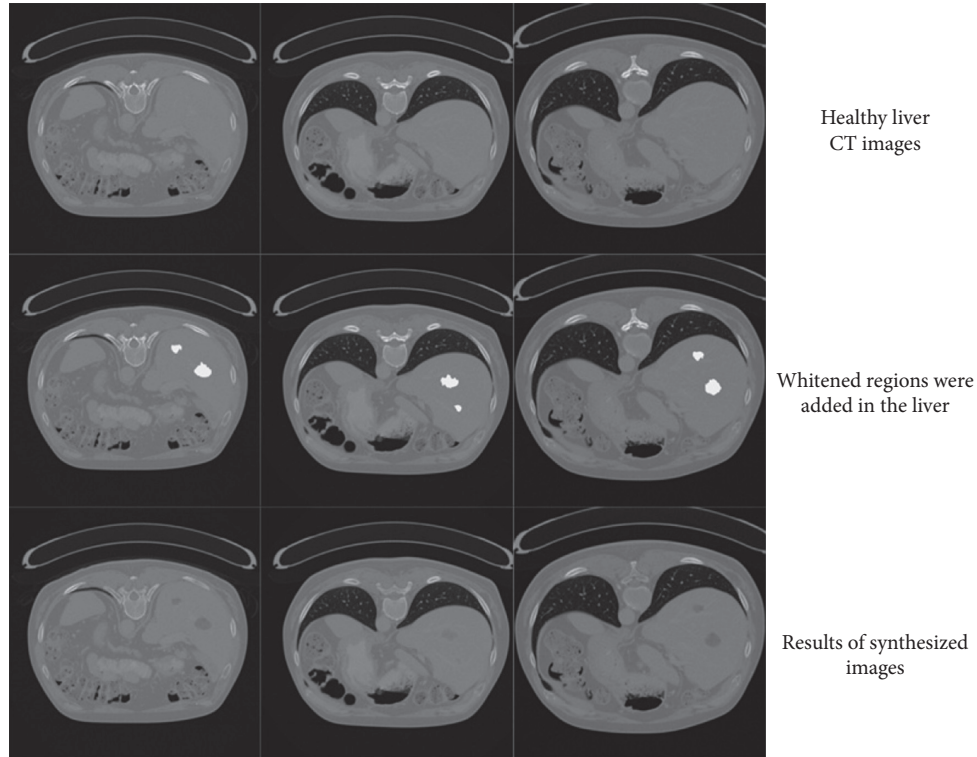


FIGURE 10: Adding tumor regions on the healthy liver CT images and synthesizing diseased liver CT images using our method.

TABLE 3: The quantitative comparison between our method and seven other approaches.

	Method							
	Atlas [17]	SR	SRF+ [18]	MOS [19]	DCAN [8]	MC-GAN [20]	ME-cGAN [21]	Our method
Mean PSNR(dB)	45.15	49.77	55.30	60.11	58.26	59.29	61.35	64.72

TABLE 4: The Turing test of our method.

	Real image (50 slices)		Synthesized image (50 slices)	
	Be judged as real images	Be judged as synthesized images	Be judged as real images	Be judged as synthesized images
Radiologist number 1	37	13	18	32
Radiologist number 2	42	8	44	6

## 5. Conclusions

In the present study, we proposed a method of synthesizing liver CT images with tumors based on mask attention generative adversarial networks. The experimental results showed that our method outperformed the other seven widely used approaches and can achieve 64.72 db mean PSNR, and the Turing test indicated that even the experienced radiologists cannot tell the differences between the synthesized images from our method and the real ones. All

these results meant that, using our method, we can build a huge medical image dataset to facilitate the diagnosis of computer-aided diagnosis and the training of deep learning.

## Data Availability

Liver CT images used in our method were from a public liver CT dataset, which is Liver Tumor Segmentation (LiTS), and the data can be obtained from <https://academictorrents.com/details/27772adef6f563a1ecc0ae19a528b956e6c803ce>.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This research was funded by the National key Research and Development Project (2018YFB2003200), National Natural Science Foundation of China (61973058), and Fundamental Research Funds for the Central Universities (N2004020).

## References

- [1] Y. Wang, L. Zhou, B. Yu et al., "3D auto-context-based locality adaptive multi-modality GANs for PET synthesis," *IEEE Transactions on Medical Imaging*, vol. 38, no. 6, pp. 1328–1339, 2019.
- [2] Y. Wang, B. Yu, L. Wang et al., "3D conditional generative adversarial networks for high-quality PET image estimation at low dose," *Neuroimage*, vol. 174, pp. 550–562, 2018. <http://www.vlfeat.org/matconvnet/pretrained>.
- [3] C. Baur, S. Albarqouni, and N. Navab, "Generating highly realistic images of skin lesions with GANs," *Lecture Notes in Computer Science*, Springer, Berlin, Germany, pp. 260–267, 2018.
- [5] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.
- [6] Q. Yang, P. Yan, Y. Zhang et al., "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.
- [7] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, and J. Paisley, "An adversarial learning approach to medical image synthesis for lesion detection," 2019, <https://arxiv.org/abs/1810.10850>.
- [8] D. Nie, R. Trullo, J. Lian et al., "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.
- [9] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. Berg, and I. Išgum, "Deep MR to CT synthesis using unpaired data," in *Proceedings of the 2017 International Workshop on Simulation and Synthesis in Medical Imaging*, Quebec City, Canada, 2017.
- [10] C.-B. Jin, H. Kim, M. Liu et al., "Deep CT to MR synthesis using paired and unpaired data," *Sensors*, vol. 19, no. 10, p. 2361, 2019.
- [11] P. Costa, A. Galdran, M. I. Meyer et al., "End-to-end adversarial retinal image synthesis," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 781–791, 2018.
- [12] P. Isola, J. Zhu, and T. Zhou, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976, Honolulu, HI, USA, July 2017.
- [13] [https://competitions.codalab.org/competitions/17094#learn\\_the\\_details-overview](https://competitions.codalab.org/competitions/17094#learn_the_details-overview).
- [14] P. Bilic, P. F. Christ, E. Vorontsov, and G. Chlebbus, "The liver tumor segmentation benchmark (LiTS)," 2018, <https://arxiv.org/abs/1901.04056>.
- [15] O. Ronneberger, P. Fischer, T. Brox, and U-Net, "U-net: convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science*, Springer, vol. 9351, pp. 234–241, Berlin, Germany, 2015.
- [16] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proceedings of the International Conference of Learning Representations (ICLR)*, Princeton, NJ, USA, May 2016.
- [17] T. Vercauteren, "Diffeomorphic demons: efficient Non-parametric Image Registration," *Neuroimage*, vol. 45, no. S61–S72, p. 1, 2009.
- [18] T. Huynh, Y. Gao, J. Kang et al., "Estimating ct image from MRI data using structured random forest and auto-context model," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 174–183, 2016.
- [19] S. Liu, E. Gibson, S. Grbic, Z. Xu, A. A. Arnaud et al., "Decompose to manipulate: manipulable object synthesis in 3D medical images with structured image decomposition," 2019, <https://arxiv.org/abs/1812.01737>.
- [20] H. Park, Y.J. Yoo, and N. Kwak, "MC-GAN: Multi-conditional generative adversarial network for image synthesis," 2018, <https://arxiv.org/abs/1805.01123>.
- [21] Y. Ren, Z. Zhu, Y. Li, and J. Lo, "Mask embedding in conditional GAN for guided synthesis of high resolution images," 2019, <https://arxiv.org/abs/1907.01710>.